Progress Report

Understanding Flight Delay

Objective

The objective of this project is two-fold. First, we aim to analyse the factors that lead to delay in passenger aircraft. As a part of this analysis, we will also investigate the effect a flight delay at a particular date and time has on subsequent delays, and under what conditions a flight delay can cause a chain reaction of delays that follows.

As the second phase of the project, we will build a model that predicts the arrival delay for a given scheduled flight. Through the analysis that we will perform in the first phase of the project, we aim to create a rich set of features that will help us predict flight delays. We also hope to attain valuable insights from the analysis, which will help us in picking the kind of prediction model that will be appropriate for our task.

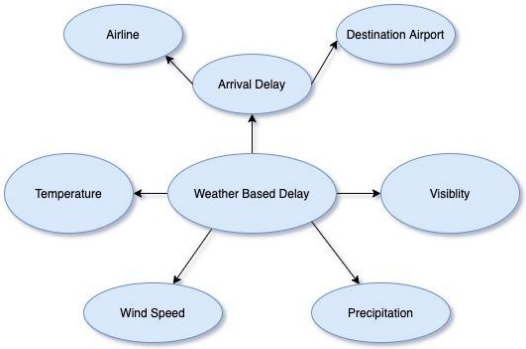
Datasets

1. Detailed Statistics Departures dataset, https://www.transtats.bts.gov/ONTIME/Departures.aspx
2. OpenSky Aircraft Database, https://opensky-network.org/datasets/metadata/
3. Iowa Environmental Mesonet https://mesonet.agron.iastate.edu/

Data Analysis

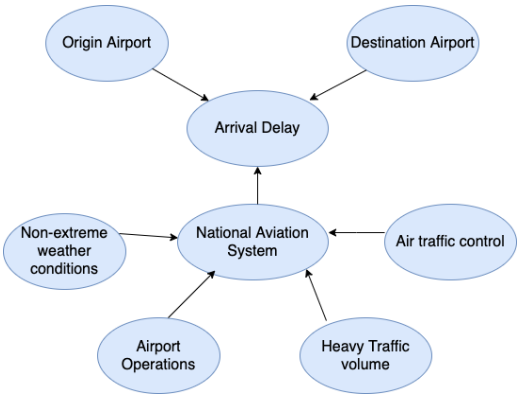
Here we tried plotting routes on the US map for different delay causes. In comparison to the lighter coloured routes, lines drawn with a stronger hue(shade) depict routes with larger percentages of delays.

Weather Delays



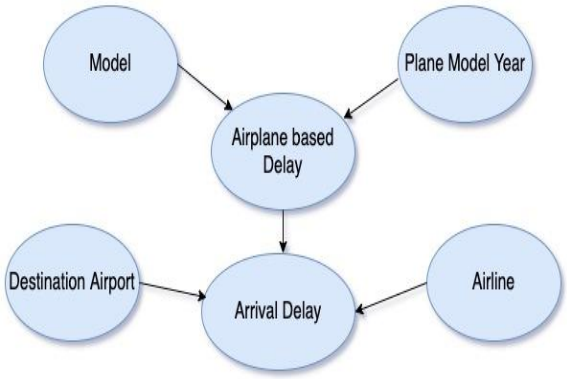
Weather delays as we can see are limited to certain routes with atypical or extreme weather conditions. According to the Federal Aviation Administration, most of the delays in winter are due to surface winds, low ceiling and low visibility, whereas during summer the majority of delays is attributable to convective weather, low ceiling and associated low visibility (Federal Aviation Administration, 2017). In our analysis, we used precipitation, wind speed, temperature and visibility as a proxy for these conditions.

National Aviation System(NAS) Delays



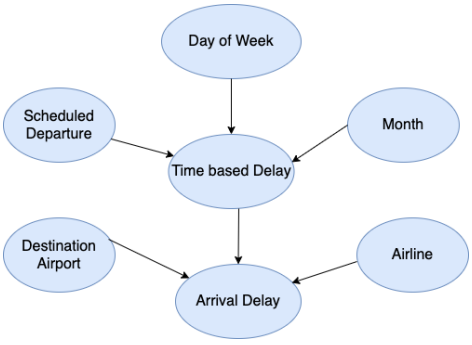
Delays and cancellations attributable to the national aviation system refer to a broad set of conditions(shown above). Since the majority component(about 45%) of NAS delays attribute to non-extreme weather conditions, resulting in similar delay routes and hotspots as compared to the weather delay graph.

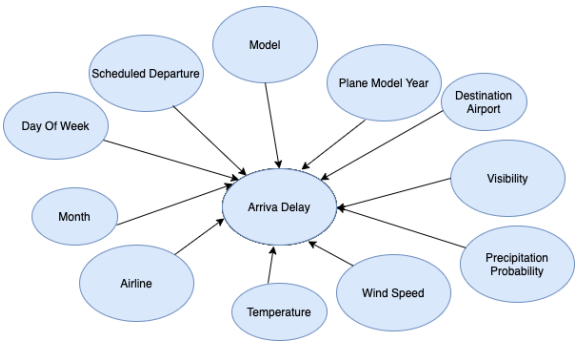
Airline and Aircraft Delays



Airline delay is more distributed across the country and is not dependent on geographical location. Upon analysis we found aircraft age and model to be important features that could affect the performance of a certain aircraft belonging to an airline.

Additional delay causes





We discovered that time-based features can also be crucial elements in forecasting flight delays after further investigation. On weekends and holidays, people prefer to travel more, causing traffic congestion and airport operations challenges, resulting in delays. We also discovered that the busiest airports had a greater likelihood of flight delays. Most of the airports that cause delays include JFK, Chicago, ATL, Miami, DNV, Austin, LA, and SF.

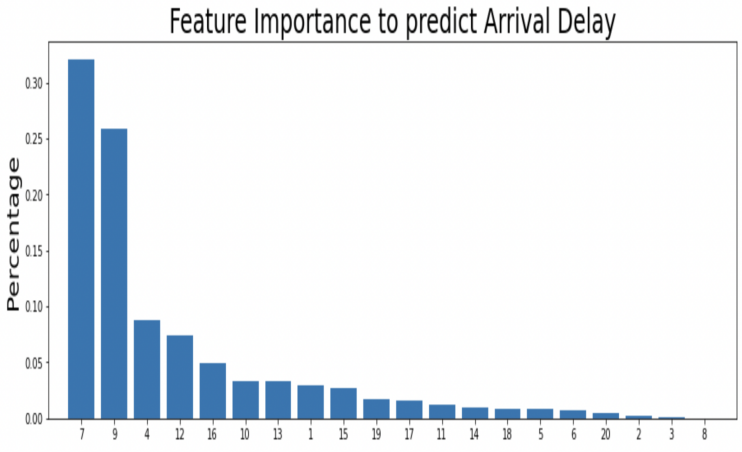
Implementation and Validation

Feature Selection

Below are the dataset features we decided based on our initial analysis by plotting the correlation between the arrival delay and other features. Data collection and cleaning took the majority of our time as Weather and Aircraft metadata was not readily available to use on the internet.

Datasets	Features	Challenges
Detailed Statistics Departures	'MONTH', 'DAY', 'DAY_OF_WEEK', 'AIRLINE', 'ORIGIN_AIRPORT', 'DESTINATION_AIRPORT', 'SCHEDULED_DEPARTURE', 'DISTANCE', 'SCHEDULED_ARRIVAL',	No challenges, data was readily available
Iowa Environmental Mesonet	'DEPARTURE_TEMPERATURE', 'DEPARTURE_WIND_SPEED', 'DEPARTURE_PRECIPITATION', 'DEPARTURE_VISIBILITY', 'ARRIVAL_TEMPERATURE', 'ARRIVAL_WIND_SPEED', 'ARRIVAL_PRECIPITATION', 'ARRIVAL_VISIBILITY',	This was not readily available. OpenWeather API was paid and there was no access to historical data. We then wrote a custom script to scrape the data from Iowa
OpenSky Network	'MANUFACTURER', 'MODEL', 'AIRCRAFT_AGE'	Not readily available. Wrote a custom script to get the desired data.

Feature Importance




```
Feature ranking:
1. SCHEDULED_DEPARTURE 0.320956
2. SCHEDULED_ARRIVAL 0.259259
3. AIRLINE 0.088341
4. DEPARTURE_PRECIPITATION 0.074216
5. ARRIVAL_PRECIPITATION 0.048809
6. DEPARTURE_TEMPERATURE 0.033440
7. DEPARTURE_VISIBILITY 0.032625
8. MONTH 0.029053
9. ARRIVAL_WIND_SPEED 0.027380
10. MODEL 0.016909
11. ARRIVAL_VISIBILITY 0.015420
12. DEPARTURE_WIND_SPEED 0.012611
13. ARRIVAL_TEMPERATURE 0.009645
14. MANUFACTURER 0.008195
15. ORIGIN_AIRPORT 0.008020
16. DESTINATION_AIRPORT 0.007151
17. AIRCRAFT_AGE 0.004596
18. DAY 0.002056
19. DAY_OF_WEEK 0.001030
20. DISTANCE 0.000291
```

Based on the features selected we used sklearn to get the importance of the overall classification. As expected, weather-based features play an important role while determining the delays across the flights. This is also supported by the analysis done by BTS which shows that almost 30% of the flights are delayed due to weather-based reasons.

Regression

We used the dataset described above to train regression models that predict the arrival delay of flights. This regression was done on a small subset of flight data, using 500000 rows for training and validation. We got a score that was greatly lower than what we expected, with our best performing model being Random Forest with an R2 Score of 0.0761.

We attribute this to the fact that our dataset was small and restricted to only one year of flight delay data(flight delays in 2015). By training our model on multiple years of flight delay data, we hope that it will be better able to understand and predict flight delays.

Model	Linear Regression	Random Forest	Neural Network(MLP)
Hyper-parameters	Default	n_estimators=100	hidden_layers=(8, 27)
R2 Score	0.0295	0.0761	0.0407

Classification

We also predicted flight delay via classification. The Federal Aviation Administration considers arrival delay greater than 15 minutes as a flight delay. Around 21.9% of the flights in our dataset were delayed according to this threshold.

When treating flight delay to be a classification problem, we got better results than with regression. Our best performing model was once again Random Forest.

Arrival Delay Threshold	Approximate % of flights on time	Accuracy
10 mins	50%	63%
15 mins	78%	77%
30 mins	90%	89%

We observe that by increasing the threshold for delay the accuracy of our model greatly improves. There is a strong correlation between the skew in our data(ratio of on-time flights vs delayed flights) and our model's accuracy. From this, we can conclude that our present model is skewed towards predicting flights as 'not delayed', and we need to identify ways to accommodate for this.

Challenges

- For training our model, we require historical weather data, as our model is trained on past flight data. The OpenWeather API did not have this historical data readily available, and hence we had to procure the same from the Iowa Environmental Mesonet dataset. As the IEM dataset is not maintained in a very user-friendly format, considerable effort was needed before we could use this data.
- When treating predicting flight delays as a classification problem, care needs to be taken as the data tends to become skewed. As a majority of flights tend not to be delayed, we can obtain an accuracy of around 80% by simply predicting every flight as 'not delayed'. Therefore, accounting for this bias when evaluating our model proved to be a challenge.
- The platform used to train our models, Google Colab, is not capable of handling the entire flight dataset. We will use a more powerful platform, such as Google Cloud Compute, to train future models.

Next steps

- So far we have trained our models using the default hyperparameters. We will perform hyper-parameter tuning using Grid CV and Randomized Grid CV.
- We have trained our model using 2015 flight data. We will now train the model using data from multiple years and try to predict flight delays for a given year using data from past years.
- So far we have been using basic scoring metrics to validate our data. We will use more sophisticated techniques like cross-validation in the future.
- The industry standard for flight delay prediction is 70-80% for classifying flight delay in buckets of 30 minutes of 1 hour. However, as more than 70% of flights are not delayed, this accuracy does not paint the correct picture of the model's performance due to the skewed nature of the data. We should take into consideration factors like precision and recall in addition to accuracy in order to fully evaluate a flight prediction model.
- Currently, we only focused on arrival delays and their relation with other features, in the later stage of the project we would also like to analyse the importance of departure delay and how it affects the overall delay in aircraft.

References

1. L. Carvalho, A. Sternberg, L. Maia Goncalves, A. Beatriz Cruz, J.A. Soares, D. Brandao, D. Carvalho, e E. Ogasawara, 2020, On the relevance of data science for flight delay research: a systematic review, Transport Reviews
2. The United States Bureau of Transportation Statistics, Understanding the Reporting of Causes of Flight Delays and Cancellations
3. Juan Jose Rebollo, Hamsa Balakrishnan, Characterization and prediction of air traffic delays
4. Borsky, S. and Unterberger, C., 2019. Bad weather and flight delays: the impact of sudden and slow-onset weather events. Economics of transportation, 18, pp.10-26.
5. Zoutendijk, M.; Mitici, M. Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem. Aerospace 2021, 8, 152
6. Yazdi, M.F., Kamel, S.R., Chabok, S.J.M. et al. Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. J Big Data 7, 106 (2020).