Forecasting Commercial Airline Flight Status Using Climate Variables

Yuge Xue, Jason Pereira, Hongkang Xu, Maxim Turkowski

Northeastern University

Abstract

Modern societies are characterized by a high degree of geographical mobility, brought by the development and diversification of transport systems. This growing demand for mobility relies heavily on transport systems. As shown throughout the Covid-19 pandemic, restricted access to key transport systems can significantly impact price levels.

One sector, which regularly experiences delays, is the air traffic sector. Between 2004 and 2017, roughly 22% of all flights within the United States were either cancelled or delayed (The Bureau of Transport, 2018). Forbes estimated that every minute delayed results in an average reduction of USD 1.42 in delayed flight ticket prices (Luttman, 2019). Obviously, delays are also inconvenient when it comes to keeping appointments or flying with travel plans.

Adverse weather conditions are an important external factor affecting delays in the air traffic system, accounting for up to 50 percent of all delayed flights pre-Covid19 (Federal Aviation Administration, 2017). With the world economy (and air traffic system) in gradual recovery, 2 more flights will again become affected under adverse weather conditions. The ability to forecast these onset events are crucial to reducing costs and improving efficiency.

Defining The Problem

The main focus of our project is to answer the following:

1. Are there existing correlations between climate variables and time-matching flight status of significance?
2. How can we effectively utilize derived trends to forecast future flight travel status based on forecasted climate variables?

Motivation

Our motivation for picking this topic stems from personal experiences with flights being unexpectedly cancelled or delayed. Nothing feels worse than finishing finals and all you can think about is enjoying a break with your friends and family and travel complications start you off on the wrong foot. It has happened multiple times where we have gone to Logan Airport on a snowy day, gone through all the hassle of passing through security, sitting at the gate, and sometimes boarding the flight just for it to get delayed or cancelled due to weather conditions. These experiences have motivated us to see if we can predict whether or not a flight will be delayed or cancelled just based on that day's weather report.

Methodology

Our initial plan was to acquire both airport data and weather data from major API providers such as AviationStack, OpenWeatherMap, etc. However, we quickly realized the drawbacks of those APIs:

1. Most APIs do not provide historical flight data that allows us to build our model
2. Most major weather API providers do not provide relative weather conditions at airports. The provided data were not applicable for aviation usage.
3. Most APIs do not provide free or reasonable price for the scale of our project
4. Some APIs take long time for queries, which significantly slow our progress.

We did further researches and quickly found two data sources that fit our purpose and available for us to use:

1. Boston Logan Airport data: Bureau of Transportation Statistics
2. Local Climatological Data: National Centers for Environmental Information

Logan Airport Departure Data

We were able to pull CSV file of departure data for five airlines (United, American, Spirit, JetBlue, Delta) from Logan Airport in 2019 from BTS. Example from the CSV:

Carrier	Date	...	Delay_carrier	Delay_weather	Delay_security	Delay_late_aircraft
B6	01/01/2019	...	0	0	0	0
B6	01/01/2019	...	0	0	0	15
B6	01/01/2019	...	129	7	0	0

Over objective was to find correlations between on-time/delayed departures with weather factors. We extracted scheduled departure times for each flight, along with the Delay_weather column, then merged the files for the five airlines into one file.

Local Climatological Data

We were able to gather CSV file of the weather conditions at Logan airport throughout 2019. Entries are gapped between around 5~60 minutes.

DATE	...	HourlyDewPointTemperature	HourlyPrecipitation	HourlyVisibility	HourlyWindGustSpeed	HourlyWindSpeed
2019-06-11	...	45	0.00	0.00	0	9
2019-11-27	...	47	T	0.12	0	5
2019-11-27	...	47	0.01	0.12	0	6

We extracted the weather factors that are most likely impact flights according the Service Blog, along with the time recorded for each entries. The weather factors included dew point temperature, predicipation, visibility, gusts, and wind speeds.

Merging Data, Model Training

Time	HourlyDewPointTemperature	HourlyPrecipitation	HourlyVisibility	HourlyWindGustSpeed	HourlyWindSpeed	Carrier	Delay_weather	Status
2019-01-01 16:18:00	29.0	0.00	9.94	0.0	23.0	B6	0	On-Time
2019-01-01 11:19:00	44.0	0.00	10.00	31.0	22.0	B6	35	Delayed
2019-01-01 06:24:00	48.0	0	10.00	0.0	15.0	B6	0	On-Time

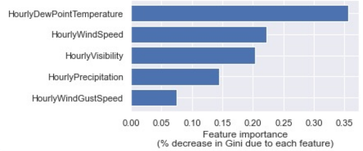
We merged the two datasets by matching departure times with the closest weather data entry from the weather data frame. From there, we classified the flights into two categories: On-time vs. Delayed. Any flights with Delay_weather equal to 0 will be classified as delayed, and vice-versa.

Initially, we decided to use the K nearest neighbor model due to its simple, intuitive, and non-parametric algorithm. We look to evaluate different K values to retrieve acceptable accuracy score. In addition to K-NearestNeighbor model, we were also looking use Random Forest Classifier. Since our target consists categorical features (on-time vs delayed), random forest classifier seems to be a good fit for this purpose. We will be splitting 70% of the data for training and 30% of the data for testing.

Results

Both KNN and RFC models produce high level of precisions in predicting the flight status with those given weather features. The closer examination of KNN model with n=5 reveals that it is has 0.99 precision level in predicting on-time flights, with 0.33 in predicting delayed flights. We determined the cause of relatively low performance in predicting delayed flights as follows:

1. The population of the delayed flights in our dataset is relatively too low comparing to on-time flights
2. The records of those weather delayed flights from BTS does not specify if the weather delay is the result of the weather at arrival airports rather than the Logan airport



Our Random Forest Classifier model produced similar accuracy at 0.988. We implemented the feature importance graph for our Random Forest Classifier. From the graph, we can tell that the Dew Point Temperature is the most dominant feature in predicting the flights. From a real-world perspective, this can be frequently observed during winter, when extreme conditions require planes to be de-iced prior to taking off, resulting in delays.

Conclusions

We believe these two models can benefit everyone taking air travels in the future. The two models in our project demonstrated high possibilities of predicting on-time flights based on weather conditions.

To answer the questions we started with, we concluded:

1. There exists fundamental correlations between climate variables to flight delays due to weather.
2. The climate variables can be used effectively in predicting on-time flights

For future usage, the potential users of this project are given the abilities to calculate the chances of on-time flights based on weather reports, thus ensuring them enjoyable on-time flights.

Any airports in the U.S where its departure data is being recorded by the BTS can be incorporated with our model to predict the flight status. Based on the similar knowledge, we have the ability to expand our research worldwide by incorporating datasets from airports outside of the U.S.

Possible Improvements

1. To increase the accuracy of the predictions on delayed flights, we may look to incorporate arrival airports weather data into the dataset.
2. Additionally, we may try to incorporate the weather data for the entire flight route to increase our accuracy for delayed flights
3. We can seek to gather datasets in addition to 2019 to gain a larger population of delayed flights
4. Port the models to other airports where weather delays are common to test the efficiency of predicting delayed flights.

References

"Bureau of Transportation Statistics." *On-Time Index Page*, https://www.transtats.bts.gov/ONTIME/

"National Centers for Environmental Information", Database, https://www.ncei.noaa.gov/products/land-based-station/local-climatological-data

Busson, Thomas. "Why Is My Flight Delayed? the 20 Main Reasons for Flight Delays." *The Service Blog*, The Service Blog, 1 Sept. 2021, https://www.getservice.com/blog/why-is-my-flight-delayed/