DS 3000 Phase 4: Forecasting Commercial Airline Flight Status Using Basic Climate Variables

Group 12: Yuge Xue, Jason Pereira, Hongkang Xu, Maxim Turkowski

Background

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 [1]. Forbes estimated that every minute delayed results in an average reduction of USD 1.42 in delayed flight ticket prices [2]. 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 [3]. 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.

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.

Model Selection

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, we are also looking to use a Random Forest Classifier, that can be used to compute and compare the feature importance of all climate variables separately. We will be splitting 70% of the data for training and 30% of the data for testing.

Data Acquisition

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 useful historical flight data. Most major weather API providers do not provide relative weather conditions at airports. The provided data was not applicable for aviation usage.
2. Most APIs are not at reasonable price for our work scale.
3. Some APIs take long time for queries, which 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

The BTS data provides on-time statistics that details the delays of each flight. Since our plan was to determine the correlation between climate variables and flight delays, we would need to gather the departure dataset since weather factors would most likely impact planes that are on-ground. Furthermore, the aviation industry has taken its worst impact last year due to the COVID-19 pandemic, which resulted a relatively small size of dataset in 2020 available for us to use. Therefore, we pulled five airlines (Jetblue, American, United, Delta, Spirit) departure data from Logan Airport in 2019, and combined then into one single file.

The Local Climatological Data provides insights on historical professional aviation weather reports. We were able to examine multiple factors such as wind/gust speeds, precip, visibility and temperature at Logan Airport. We pulled the 2019 data which contains weather reports for roughly every five minutes.

References

[1] Bureau of Transport Statistics, 2018. 'On-Time performance- flight delay at a Glance', United States Department of transportation. https://www.transtats.bts.gov/.

[2] Federal Aviation Administration, 2017. NextGen Weather. United States Department of Transport. https://www.faa.gov/nextgen/programs/weather/.

[3] Luttmann, Alexander. "Are Passengers Compensated for Incurring an Airport Layover? Estimating the Value of Layover Time in the U.S. Airline Industry." Economics of Transportation, vol. 17, 2019, pp. 1–13., https://doi.org/https://doi.org/10.1016/j.ecotra.2018.11.002.

Merging Datasets

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.

Results

Both KNN and RFC models produce high level of precisions in predicting 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. By visualizing the feature importance for our RFC model, we could tell that the Dew Point Temperature was the most dominant feature in predicting the flights. From a real-world perspective, this can be frequently observed during winter, when extreme weather conditions required planes to be de-iced prior to taking off, resulting in delays.

Conclusion

We believe these the the two trained models used in this project (KNN and Random Forest Regressor) can be applied to our everyday lives. They demonstrate the possibilities of predicting on-time flights based on basic climate variables. The accuracy of these two models show, respectively, that they can accurately predict flight status (on time/ delayed) and that the climate features chosen are relatively significant and independent. Nevertheless, this project offers a perspective into using statistical models to predicting air travel flight status given basic climate variables, which might be useful for travellers planning their flight schedule. With more people catching their flight on time, we can expect a considerable reduction within the recurring loss of additional economic resources due to air traffic.