Flight delay prediction and exploration in the United States

Team 127 - Aviators, CSE 6242, OMSA, Georgia Institute of Technology

Aaron Joseph Mathew, Ayush Subedi, Harshini Prabakaran, Hira Laghari, Saumya Shekhar

1 INTRODUCTION & MOTIVATION

The aviation sector amounts to USD 2.7 trillion to the gross domestic product (GDP) and 65.5 million jobs, which is comparable to the GDP of the United Kingdom [ICAO 2019]. Air travel has become more widespread due to the low costs of travel as well as increased safety, with roughly 100,000 flights taking off and landing daily [Myers 2016]. According to FlightAware, one in five flights have been delayed in 2022 in the United States, and a little over 3% of the flights have been canceled [Murphy 2022].

Airplanes are one of the most common modes of transportation when it comes to traveling long distances, and most airlines run on a very tight schedule. The more flights in the air, the more money the airline makes, thus it's quite important to be landing and departing at the right time. However, even despite the best efforts, it is not always possible to ensure that. There are multiple factors due to which a flight can get delayed, and this can cause a domino effect, causing subsequent flights to also be delayed. These are sometimes referred to as cascading delays.

There is an economic cost to flight delay as well and at the same time causes inconvenience to passengers. Flight delays not only irritate air passengers and disrupt their schedules but also cause a decrease in efficiency and an increase in capital costs.

1.1 Problem Definition

The objective function of this project is to find the relevant data which can give us insights into the leading reasons for flight delays and use this to predict upcoming delays.

We have developed several robust machine learning models that can help predict delays, thereby ensuring that customers can mitigate the risks associated with it. Furthermore, we have developed a tool that allows users to interact with the model on the web.

1.2 Literature Survey

Due to the high economic impact of flight delays, with costs of delays in the United States in 2007 estimated to be around USD 28.9 billion [Ball et al. 2010], and the negative impact of passenger sentiment towards airlines and airport services in the event of a delay or cancellation [Song et al. 2020], ample research has been done to identify the reason for delays as well as how to lessen them.

Several papers investigate flight delays using machine learning algorithms; [Chakrabarty 2019] used a data mining approach to predict flight arrival delays for American Airlines by using a gradient boosting classifier, which achieved an accuracy of 85.73% to predict delays. Others used causal machine learning techniques to construct the network representing causal relationships between airport variables and delay incidents, which then helped in creating a predictive model to predict flight delays [Truong 2021]. Another paper studied spatial, temporal, and extrinsic features to predict flight delays in domestic flights in China using a random forests model [Li and Jing 2022].

It is also important to investigate the various reasons why delays may occur. [Fleurquin et al. 2013] studied technical, operational, and meteorological issues propagating delays across the US airport network. The paper also establishes passenger and crew connectivity as the most relevant factor for delay spreading. We also intend to study cascading failures, i.e., delays in one airport create delays in others, for which the aforementioned study might prove to be useful. There are also many direct and indirect factors that might be responsible for flight delays, and one paper utilized LSTM-AM (long short-term memory network) mechanisms to predict this delay [Wang et al. 2022].

Adverse weather events also have a profound impact on flight delays and cancellations and being able to accurately predict these events can aid in better scheduling of flights. [Belcastro et al. 2016] propose

classification-based machine learning techniques to predict the delay of a scheduled flight due to weather conditions. It takes into consideration the flight information and the weather conditions at the origin and destination. [Borsky and Unterberger 2019] also studied the impact of bad weather by categorizing weather conditions and using time-series modeling, difference-in-difference frameworks, and Prais-Winsten estimators to predict the extent of delays caused by certain weather conditions.

Another paper used stochastic models of airline networks to identify passenger-centric metrics to measure on-time performance and improve capacity at airports which cause bottlenecks over the entire network [Arıkan et al. 2013]. It will help to provide an idea of the existing approaches that airlines take to reduce delays. Similarly, [Lambelho et al. 2020] assessed the effectiveness of strategic flight scheduling using machine learning on reducing flight delays. Notably, [Yimga 2021] found that with reduced air travel during the pandemic, flights were departing and arriving with less delay despite the downsizing in the airline industry and additional delay time for COVID-19 protocols.

2 PROPOSED METHODOLOGY

2.1 Intuition and Innovations

Some limitations in the existing research involve using smaller datasets over a shorter period of time, as well as using traditional methods to study flight delays which don't account for the complete treatment of noise when dealing with large volumes of data ([Yazdi et al. 2020]). Our dataset has significantly more features and more data points, which allows us to explore outliers and apply effective feature selection strategies. We have experimented with several effective feature selection techniques, principal component analysis, and predictive classifiers built on Catboost, Neural Nets, etc. to come up with better results compared to results in the literature we have reviewed. To measure the success of our models, we have used validation metrics such as accuracy, precision, recall, and F1 score on the testing dataset.

Additionally in this project, our aim is to use a combination of the techniques outlined, as well as offer an interactive dashboard that the end user, mainly passengers, can use to determine whether a flight has a possibility of delay. As mentioned previously, flight delays and cancellations are responsible for huge economic losses, and these losses are borne both by passengers and airlines, with [Belcastro et al. 2016] studying this effect using a generalized multilinear 3LS regression model in an attempt to quantify the cascading impact of air delays. The advantage we have compared to [Belcastro et al. 2016] is the large volume of the dataset that also allows us to identify if the delays have gotten worse over time [Kim and Park 2021].

Our goal with this project is to provide a better classifier that will help passengers plan trips more efficiently. We can measure this by calculating costs saved when passengers choose routes and airlines with a lower probability of delay. Some risks associated with this project are that even with an efficient delay prediction model, the accuracy of delay estimation could be affected majorly by external factors like uncontrollable weather conditions, macroeconomic pressures, changes in consumer behavior, etc. Conversely, the payoffs are that by quantifying and predicting delays in complex networks, we can help airlines achieve maximum utilization and reduce customer dissatisfaction.

We have made use of Machine Learning Models based on Classification Algorithms. Based on the Literature references used in our research, Random Forest(RF) has been proven to be the most accurate among other classification algorithms. The good performance of RF comes from the fact that it uses distributed intelligence, it creates decision trees based on the randomly picked feature from the list. It is an ensemble learner of multiple Decision Trees DT and the final result is calculated based on the results from multiple DTs.

In this project, random forest models are experimented with different leaf sizes and bagging of different sizes for training the model. Further, instead of making the feature selection random for the decision trees, it will make use of a custom function that would take into consideration the weight and number of occurrences of a feature. While designing the function, the performance aspect will be kept in mind to avoid unnecessary load and computation.

Training the model with a huge data load is always a challenge. To tackle such issues, the approach plans to implement MapReduce Algorithms to divide the datasets into multiple small datasets. The decision trees of RF will be trained on the different datasets.

The approach will further enhance the algorithms using Boosting Techniques. For boosting, the technique

we will be using is Cat-boost, an open-source gradient enhancement algorithm developed by Yandex Company. Cat-boost works really well with decision Trees based on Categorical Data.

The results from exploratory data analysis, presented in GCP Looker Studio, will be embedded in a Flask app. The app will also be used for model-serving purposes. Users can interact with the delay prediction models, and provide input variables. The project is currently hosted using Heroku at https://flight-delays-team-127.herokuapp.com/

2.2 Description of Approaches

2.2.1 Data Exploration and Architecture Design.
As the goal of this project is to predict future delays and cancellations, we required a detailed dataset that had ample information about delays, their causes, as well as the magnitude of delays. Our dataset was collected from the Bureau of Transportation Statistics, provided by the US Department of Transportation and it describes flight data for various airlines in the United States. The dataset consists of 109 columns, and more than 200 million rows, with data spanning the years 1988 to 2021.

The size of the data allows for detailed analysis, and some of the columns we intend to use include:

- **IATA CODE Reporting Airline**: The code assigned by IATA to identify a carrier
- **OriginAirportID**: Airport ID for the origin airport
- **DestAirportID:** Airport ID for the destination airport
- **DepDelay**: Departure delay in minutes
- **ArrDelay**: Arrival delay in minutes
- **CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay**: Various causes of delays, with the amount of delay given in minutes

To accommodate to the large size of our dataset, and develop insights and prediction models, we have built the following data pipeline. The tools we are primarily using are PySpark, Google BigQuery, Pandas, Scikit-learn and Flask.

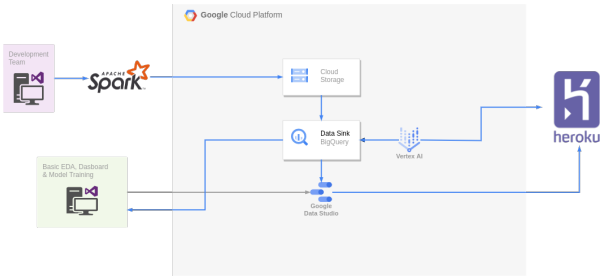


Figure 1: Architectural Diagram

2.2.2 Data Cleaning.
The initial data set on air delays consisted of 109 columns (both categorical and continuous). Many of these columns consisted of missing values without any explainability and were removed as a part of the data cleaning process. Additionally, some of the columns that were out of the scope of the project were also removed. Also, using AirportID's, we also added geographical information such as airport latitude, longitude, city, and state for the purposes of map-based visualizations on Tableau.

2.2.3 EDA.
The initial results from exploratory data analysis showed a high correlation between various types of delay. Weather delays were significantly affected by the season of travel, while arrival and departure delays were highly correlated with the origin and destination of the air route.

2.2.4 Feature Selection.
The features that are important for delay prediction were identified using the:

The results were obtained by testing on samples from every year between 2014-2022, to eliminate any selection bias and verify if there were any major changes in trends across the years. The dataset is now reduced to 12 feature variables describing the Origin, Destination, Season, Distance, Airline carrier, and Flight number, captures more than 90% of the variance in Delays, and will be used further for delay prediction

2.2.5 Google Cloud Platform.
Google Cloud Platform was used to store the data in BigQuery, which is a serverless database. Additionally, BQML, which is the BigQuery ML offering from GCP was used to develop BigQuery Models, based on the data which is prepared in the database. Furthermore, the Vertex AI platform was used to host the models. Vertex

AI is the MLOps offering from GCP, which allows for an end-to-end ML lifecycle.

2.2.6 *Dashboarding.*

The results of EDA have been reported on GCP Looker Studio. One of the key advantages of Looker Studio is that it makes use of BigQuery BI Engine to query the results, which allows for faster load time.

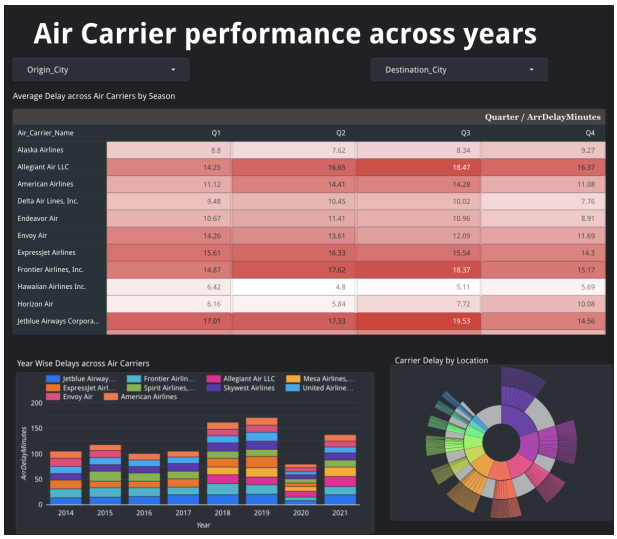


Figure 2: GCP Looker Studio Dashboard

- Flight delays are highest during the Summer months, majorly driven by Air Carrier, Logistics related delays due to the high demand and inclement weather conditions like thunderstorms
- Among the popular airlines, JetBlue and Express-Jet Airline have the highest average delay while Delta and Southwest Airlines have the least delays
- Among the popular states in US, New Jersey has the highest average flight delay of 17 minutes, while Hawaii has the least of 7 minutes

2.2.7 *Modeling.*

We built several machine-learning models to predict delays. We feature-engineered a new column called **Delay** to store our binary predictor variable. If a flight arrives at its destination fifteen minutes later than the scheduled arrival time, we consider the flight to be delayed.

Similarly, we have built two versions of the models for each of the machine learning algorithms. The first version does not use Departure Delay (difference in minutes between scheduled and actual departure time) as one of its features. The second version uses Departure Delay, and as a result, has higher accuracy in all experiments below.

The model serving page on our website has an optional field called Departure Delay in Minutes. If the user provides this information, the backend (flask app) will utilize the first version of the model to provide a prediction to the user.

2.2.8 *Model Serving.*

The model is served from the site **https://flight-delays-team-127.herokuapp.com/model_serving**

The screenshot below demonstrates delayed arrival. Both delayed and on-time arrival can be tested live on the website. The Initial Transformation tables in the screenshots constitute the predictor variables used to train the models. The model served here is based on the Random Forest models (this is explored in the experimentation section below).

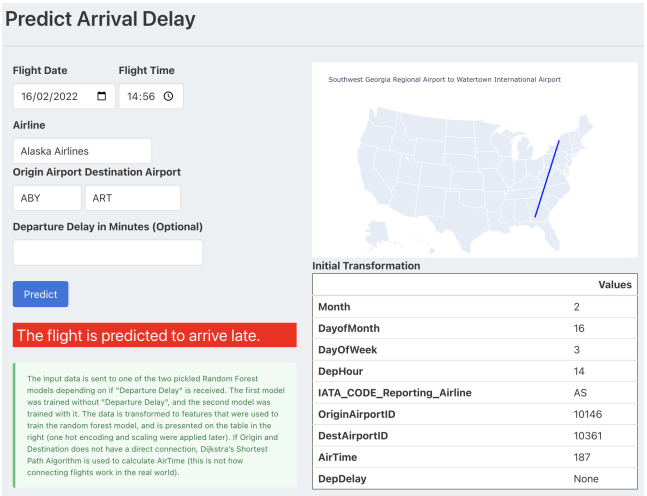


Figure 3: Prediction tool predicting a delay

2.2.9 *Model Registry - Vertex AI.*

Additionally, models that were trained in GCP BQML, were stored in Vertex AI, which is a GCP MLOPs offering. Models have been stored in Model Registry, which can be used for model inferencing, in both batch and online.

2.3 Experiments, Evaluation

2.3.1 *Experiment 1: Data re-sampling.*

We experimented if the delay factors in the original dataset that should be used as training factors. Some of

the delay factors were: CarrierDelay, NASDelay, WeatherDelay, SecurityDelay, etc. While using these parameters resulted in models with very good accuracy scores, a normal passenger is not capable of acquiring this information before a flight. Additionally, we would also be feeding in factors that are highly correlated with our outcome variables in the model. Therefore, we decided that we would not be using these delay parameters.

2.3.2 *Experiment 2: Feature selection.*

We experimented with several feature selection techniques:

- **Analysis of Variance (ANOVA)** - Features with high F-statistic score and p-value below 0.05 were selected
- **LASSO Regression** - Data was sampled across years and fitted to the LASSO model to identify features with significant coefficients
- **Stepwise OLS Regression** - To validate the results of the above two methods, we ran a Step-forward OLS regression model with adjusted R^2 as a performance metric.

2.3.3 *Experiment 3: SMOTE and Controlled Selection of Classes.*

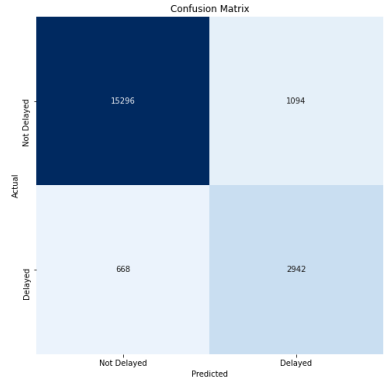
The original dataset has more data points associated with flights that arrived on time, compared to flights that were delayed. The class imbalance is an issue here. We used SMOTE to overcome this issue.

Additionally, one of the failed experiments was splitting the dataset into training and testing sets after SMOTE. Although our accuracy scores were great, splitting after SMOTE would leak training data into testing data. We fixed this issue in the subsequent iterations of model development.

Apart from SMOTE, we also experimented with a controlled selection of classes. However, we opted for SMOTE because we wanted the test data to be as close to real-world data as possible, and controlled sampling would not allow that to happen.

2.3.4 *Experiment 5: CatBoost.*

CatBoost is known to provide great results with default categorical parameters, and we experimented if this algorithm would be a better fit for our use case. The accuracy is 0.74 (without Departure Delay), and 0.95 (with Departure Delay).



Classification Report:

	precision	recall	f1-score	support
Not Delayed	0.96	0.93	0.95	16390
Delayed	0.73	0.81	0.77	3610
accuracy			0.91	20000
macro avg	0.84	0.87	0.86	20000
weighted avg	0.92	0.91	0.91	20000

Figure 4: CatBoost with Departure Delay

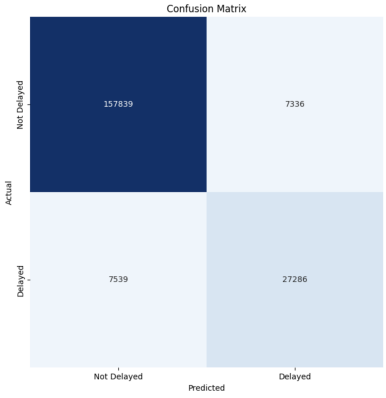
2.3.5 *Experiment 6: One hot encoding and scaling.*

Since most of our variables were categorical (or categorical variables disguised as numeric variables), we experimented with one-hot encoding and scaling. Standard scaling was used for Fight Distance.

All our models (apart from CatBoost) use this.

2.3.6 *Experiment 7: Random Forest Classifier.*

The model used in the model serving page of our website is based on the Random Forest model, and its evaluation scores are presented below. The accuracy is 0.72 (without Departure Delay), and 0.93 (with Departure Delay).



Classification Report:

	precision	recall	f1-score	support
Not Delayed	0.95	0.96	0.95	165175
Delayed	0.79	0.78	0.79	34825
accuracy			0.93	200000
macro avg	0.87	0.87	0.87	200000
weighted avg	0.93	0.93	0.93	200000

Figure 5: Random Forest with Departure Delay

2.3.7 *Experiment 8: One Airport and One Year.*

We replicated the Random Forest experiment for one airport only (ATL), and separately for one year (2015). The results were pretty consistent to previous model. The results might be different for an obscure airport, but we were interested to experiment if one of the busiest airports encapsulated the results we previously saw. However, the delay trends for Covid years are a little different based on our EDA.

2.3.8 *Experiment 9: Algorithms that worked: Boosting, Bagging, and Neural Nets.*

All the boosting algorithms (CatBoost, XGBoost, AdaBoost) resulted in great accuracy scores. Similarly, bagging algorithms such as Random Forest also performed well. Neural Nets also resulted in great accuracy scores. The accuracy scores of Neural Net Classifier are 0.80 (without Departure Delay) and 0.89 (with Departure Delay), and for AdaBoost are 0.81 (without Departure Delay) and 0.92 (with Departure Delay).

2.3.9 *Experiment 10: Algorithms that did not work: Naive Bayes and Quadratic Decision Boundary.*

Algorithms such as Naive Bayes and Quadratic Decision Boundary resulted in poor scores. The accuracy scores for Naive Bayes are 0.26 (without Departure Delay) and 0.28 (with Departure Delay). The accuracy scores for Quadratic Decision Boundary Classifier are 0.55 (without Departure Delay) and 0.64 (with Departure Delay).

2.3.10 *Experiment 11: GCP BigQuery ML & Vertex AI.*

GCP BigQuery was used as a Data Warehouse for our project. Herein, GCP BigQuery served as the backend engine for the visualization tool - GCP Looker Studio and at the same time, BigQuery also offers BigQuery ML, which is a managed service for developing models based on BigQuery data. Two models were developed, Deep Neural Network Classifier and Wide and Deep Neural Network Classifier. The features that were used in BigQuery ML differ from the previous experiments, this is due to the addition of data points that were done in BigQuery. The model performance is as follows:

Deep Neural Network

(1) With Departure Delay - 93.9%
(2) Without Departure Delay - 84.06%

Wide and Deep Neural Network

With Departure Delay - 91.66%
Without Departure Delay - 73.8%

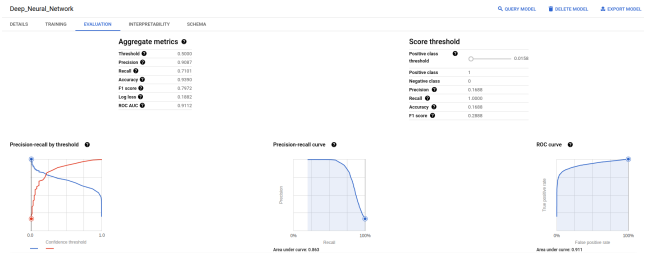


Figure 6: BigQuery ML Evaluation Results - Deep Neural Network

3 CONCLUSION AND DISCUSSION

This project has highlighted the use of machine learning systems to predict the likelihood of delay. Additionally, the Google Cloud Platform was used for Data Warehousing, Model Development, Model Registry, and Deployment. At the same time, an independent app was created using Flask and deployed over Heroku, which has all of the functionality that we had originally planned for. Furthermore, the Dashboard that was created in GCP Looker Studio has been embedded into the Flask App to bring about a unified view. Looker Studio is powered by BigQuery BI Engine which allows for faster load time.

In terms of our key findings, flight delays are the highest during the summer season. Additionally, using explainable AI, two major factors that can predict Airline delay are Departure delay and Airlines Carrier Name.

In terms of how our models performed compared to those of our peers, most papers using a Random Forests algorithm achieved better results, within the 90% range, however, all the papers we studied used delay factors to determine delay. What is unique about our approach is we provided the option to input departure delay, but also predicted delay without using any delay factors as that will greatly affect model accuracy. We also had our website surveyed by friends and family to ensure that the website is accessible and easy to navigate for users.

The analysis of cascading delays has been omitted due to time constraints, and due to the high level of complexity associated with such analyses.

3.1 Team Member Contributions

All members of the team have contributed equally to each task and deliverable.

REFERENCES

Mazhar Arıkan, Vinayak Deshpande, and Milind Sohoni. 2013. Building Reliable Air-Travel Infrastructure Using Empirical Data and Stochastic Models of Airline Networks. *Operations Research* 61, 1 (2013), 45–64. http://www.jstor.org/stable/23482073

Michael Ball, C. Barnhart, Martin Dresner, Mark Hansen, K Neels, Everett Odoni, A.R.and Peterson, Lance Sherry, Antonio Trani, and Bo Zou. 2010. Total delay impact study : a comprehensive assessment of the costs and impacts of flight delay in the United States. *U.S. DEPARTMENT OF TRANSPORTATION* (2010). https://rosap.ntl.bts.gov/view/dot/6234

Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. 2016. Using Scalable Data Mining for Predicting Flight Delays. *ACM Trans. Intell. Syst. Technol.* 8, 1, Article 5 (jul 2016), 20 pages. https://doi.org/10.1145/2888402

Stefan Borsky and Christian Unterberger. 2019. Bad weather and flight delays: The impact of sudden and slow onset weather events. *Economics of Transportation* 18 (2019), 10–26. https://doi.org/10.1016/j.ecotra.2019.02.002

Navoneel Chakrabarty. 2019. A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines. In *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*. 102–107. https://doi.org/10.1109/IEMECONX.2019.8876970

Pablo Fleurquin, J. Ramasco, and V Eguiluz. 2013. Systemic delay propagation in the US airport network. *Scientific Reports* 3 (2013). https://doi.org/10.1038/srep01159

ICAO. 2019. Aviation Benifits Report. https://www.icao.int/sustainability/Documents/AVIATION-BENEFITS-2019-web.pdf

Myeonghyeon Kim and Sunwook Park. 2021. Airport and route classification by modelling flight delay propagation. *Journal of Air Transport Management* 93 (2021), 102045. https://doi.org/10.1016/j.jairtraman.2021.102045

Miguel Lambelho, Mihaela Mitici, Simon Pickup, and Alan Marsden. 2020. Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *Journal of Air Transport Management* 82 (2020), 101737. https://doi.org/10.1016/j.jairtraman.2019.101737

Qiang Li and Ranzhe Jing. 2022. Flight delay prediction from spatial and temporal perspective. *Expert Systems with Applications* 205 (2022), 117662. https://doi.org/10.1016/j.eswa.2022.117662

Heather Murphy. 2022. Understanding the summer air travel mess. https://www.nytimes.com/2022/07/01/travel/summer-travel-flight-delays-cancellations.html#:~:text=So%20far%20in%202022%2C%20an,116%2C000%20flights%20have%20been%20canceled

Joe Myers. 2016. This visualization shows you 24 Hours of global air traffic – in just 4 seconds. https://www.weforum.org/agenda/2016/07/this-visualization-shows-you-24-hours-of-global-air-traffic-in-just-4-seconds/

Cen Song, Jingquan Guo, and Jun Zhuang. 2020. Analyzing passengers' emotions following flight delays- a 2011–2019 case study on SKYTRAX comments. *Journal of Air Transport Management* 89 (2020), 101903. https://doi.org/10.1016/j.jairtraman.2020.101903

Dothang Truong. 2021. Using causal machine learning for predicting the risk of flight delays in air transportation. *Journal of Air Transport Management* 91 (2021), 101993. https://doi.org/10.1016/j.jairtraman.2020.101993

Fujun Wang, Jun Bi, Dongfan Xie, and Xiaomei Zhao. 2022. Flight delay forecasting and analysis of direct and indirect factors. *IET Intelligent Transport Systems* 16, 7 (2022), 890–907. https://doi.org/10.1049/itr2.12183

M.F. Yazdi, S.R. Kamel, and S.J.M. et al. Chabok. 2020. Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data* 9 (2020). https://doi.org/10.1186/s40537-020-00380-z

Jules Yimga. 2021. The airline on-time performance impacts of the COVID-19 pandemic. *Transportation Research Interdisciplinary Perspectives* 10 (2021), 100386. https://doi.org/10.1016/j.trip.2021.100386