Domestic U.S. Flight Delay Prediction

Group 36 - Yigithan Akercan, Shaambhav Dave, Zachary Dearman,
Ronith Gonsalves, Anish Jaisinghani, & Krishna Maran

Introduction

In the fast world of air travel, flight delays and cancellations are a common frustration for passengers, impacting operational efficiency, economic production, and customer satisfaction. Understanding the root of these delays and being able to predict them consistently is critically important in the aviation industry, for airlines, airports, and travelers alike.

The motivation behind developing a predictive flight delay algorithm is to create a reliable model for airlines to optimize operations by adjusting schedules and reallocating resources, saving money and improving customer satisfaction. Further, our model and associated interactive visual is aimed at travelers to better inform them of potential delays they may experience, minimizing unexpected disruptions and allowing for better planning.

Problem Definition

Using historical flight data, we aim to develop a machine learning algorithm and interactive interface capable of predicting flight delays in the domestic United States. As previously mentioned, this tool will help individuals make informed decisions about how and when they travel, and provide airlines with the information needed to properly allocate time and resources.

Literature Survey

Flight delays and cancellations, often caused by technical defects and delayed entry[1], are frustrating and costly to passengers and airlines. We aim to build a tool which considers airline, airport, and time to accurately predict whether a given flight will be delayed[5].

Current flight delay prediction models primarily use temporal, network, operational, and weather data from government agencies to train and run their algorithms, combining machine learning and statistical analysis to predict flight delays[4][8][9].Specifically, the comparative efficiency of models like RNN, LSTM, and random forest are useful in analyzing delay patterns, with the latter emerging as most accurate, effectively addressing overfitting problems and providing a robust analysis framework for improving prediction capabilities[12]. An examination of data science approaches to flight delay analysis offers a classification of methods and evaluates solutions, stressing the need to refine methodologies to boost model precision in the ever-evolving context of air transport[7][15]. However, there are three major concerns when dealing with this: root delay, delay propagation, and cancellation, and many authors choose to only focus on one of these concerns to model. Further, these models often use outdated data[2], too small of a dataset[8], or are too computationally expensive[9], whereas our analysis will utilize a large, relevant dataset at no monetary cost. In the past 5 years, the operation and demand for commercial airline travel has fluctuated dramatically due to COVID-19. Models made with current data have had to account for these rapid growths and declines in the number of flights[4].

Flight delays impact customers, airlines, and airports, and the financial cost to the US economy exceeds $30 billion, and a predictive model will help airlines mitigate delays and airports to account for delays in their flight schedule[2]. Further, successfully predicting airline delays can significantly improve flight travel efficiency and passenger satisfaction, leading to fewer disruptions, higher cost savings, and better resource allocation. To measure the impact, tracking metrics can compare predicted and actual delays while user studies can monitor customer satisfaction scores.

Using all 3 aspects of flight delays in the same model is a risk as it adds complexity, which can make it harder to interpret and explain, and there is a greater risk of overfitting when adjusting to new data. Payoffs of this model are increased accuracy, as it considers all types of delays at once,

and satisfaction, gaining popularity as a result of minimized unexpected delays. The versatility allows the model to handle sudden changes in weather, global events, or schedule.

Our novel flight delay prediction algorithm synthesizes root delay, delay propagation, and cancellations into a single, dynamic model. Unlike conventional approaches that rely solely on historical data[16][18], our method embraces adaptive modeling, updating with real-time information to accurately forecast delays, even when unprecedented events occur. By incorporating a diverse range of data sources, including weather updates and alongside traditional air traffic data, our algorithm will create a comprehensive and innovative approach to anticipating flight delays. This integration of disparate prediction methods into a unified model will vastly improve how flight delays are managed, offering airlines and passengers a more reliable, responsive tool when traveling.

To better develop and enhance our model, we considered other existing models, and their methodology. A random forest model based on spatial and temporal data considered multiple factors into one model, but applied only to Chinese domestic flights[11]. One model used multiple linear regressions with departure delays and distances to predict arrival delays, proving useful as a quantitative approach but doesn't account for events that may occur[6][14]. One model of the JFK Airport compared seven algorithms and provided indicators of success, but these algorithms were assessed over only a one-year period[13]. Using a Bayesian Network with a delay tree framework, one model worked to identify how flight delays propagate through a network of airlines, focusing on factors such as connecting flights, but this approach was highly computationally expensive and had potential for overfitting [17]. Many models do not consider canceled flights, or focus solely on this metric, but in considering both delayed and canceled flights, we can provide critical information to users[10].Lastly, a model utilizing parallel algorithms on a cloud platform using scalable data mining is considered, but this approach was heavily reliant on weather as the primary delay factor[3][16].

Proposed Method

Our visualization is a web page that utilizes D3 to load and display our flight delay prediction algorithm. There are two ways to use our tool, each serving its own purpose. The first way is to select an airport, airline, and month to display the predicted delay length and cancellation probability. This is for users who have already booked their flight and want to get an accurate prediction for their flight. After pressing submit, the D3 script gets the user's input options, loads our algorithm data and displays the appropriate results. The predicted delay time and cancellation probability are displayed in a box below the dropdown options for ease of use. If the user's input options are out of the scope of our algorithm, "Not Available" is displayed.

The second way to use our tool is for users want to fly out of a city but have not chosen a particular airline yet. Using D3, we created a projection of the US and overlaid pins on top of the map with the 63 busiest airports in the US. When the user hovers over an airport an event listener is triggered that gets all airlines that service the selected airport. The script then filters the 5 airlines with the lowest expected delays and displays them in increasing order in bar chart form. This is so users can pick the airline that will have the lowest delay time when booking flight tickets.

In terms of the algorithm, for the data for visualization of delay time predictions according to airline, departure location, and month of travel, we have implemented a Random Forest Regressor algorithm to predict delay times of future data. When cross-validated with averages, this algorithm yielded a consistent prediction. One other approach of ours was to use a Gradient Boosting Regressor to predict continuous delay time values. However, when cross-validated, the Gradient Boosting Regressor algorithm didn't give sufficiently consistent results because it wasn't as accurately handling negative values representing early departures as the Random Forest Regressor did.

For the flight cancellation probabilities, we used directly the averaging method for percentages because of imbalanced or 0 positive values for some of the groups as cancellations occur much rarer compared to delays.

Experiments & Evaluation

Upon our midterm check, we wanted to come up with higher accuracy scores for the whole dataset. We were wondering what features could lead to better accuracy. In addition to that, our previous model had classified delays as 1 or more minutes. However, the FAA defines a delay as a flight arriving at its scheduled destination 15 or more minutes late. Therefore we were wondering if we could achieve high accuracy scores by labeling flights 'delayed' for delay times of 15 minutes so that it would actually be useful to predict delays in its official terms.

After running our models numerous times on different features, and looking at the feature importances, we came up with the most influential features being the airline and the departure location. For better visualization, we also interpreted months of travel in our test models to check if it would negatively affect accuracy. So, Figure 1 below represents the machine learning algorithms used to check accuracy scores on train data (80%) and test data (20%) for 2 features being airline and the departure location, where if it's labeled below as 3 features, the third feature being the month of travel.

	Training Accuracy	Test Accuracy
Gradient Boosting Classifier (2 Features)	0.8236	0.8228
Gradient Boosting Classifier (3 Features)	0.8236	0.8228
Random Forest Classifier (2 Features)	0.8237	0.8227
Random Forest Classifier (3 Features)	0.8250	0.8206

Figure 1: Accuracy scores for testing and training datasets.

Even though the Gradient Boosting Classifier and Random Forest Classifier seemed to perform similarly, the Gradient Boosting Classifier had issues with classes that had no predictions. Hence, when compared to Random Forest Classifier, Random Forest achieved a better job in terms of precision, and therefore we also used Random Forest in our second interactive visualization which presents the predicted delay times generated by Random Forest Regressor grouped by airline, departure location, and month of travel.

For the graph part of our visualization, our goal was to show users, at a glance, the reliability of each airline to inform customers of delays before booking flights. We experimented with several methods to predict this reliability. First, we tried extrapolating from current years but found that year-to-year there weren't many significant changes in flight delays, aside from a roughly year-long period from 2020 to 2021 where average flight delay decreased, likely due to COVID and reduced flights. Additionally, there wasn't a strong correlation between time and average departure delay, as shown in Figure 2.

For this reason, we moved away from our original idea, which was to use a regression model along with this time series data to make predictions into the future. Because, over time, the average departure delay didn't change significantly, we decided that it would make sense to take an average over all data points as a predictor for future delays. However, we still had to decide which average we would use as a predictor. Initially, we decided to group by airline, origin city, and destination

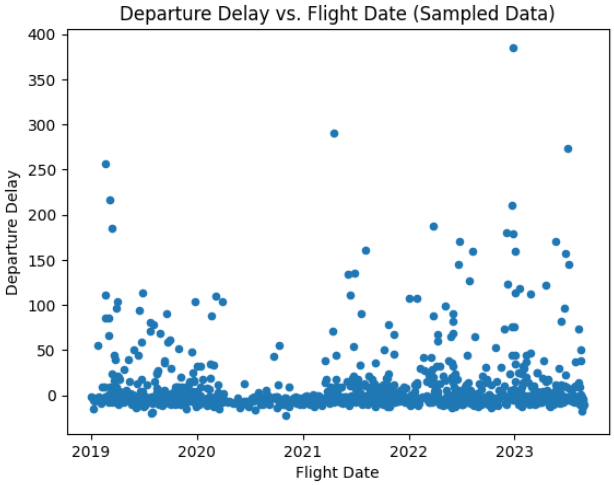

Figure 2: Random sample from data showing no clear relation between date and departure delay.

city and take the average over delays within these groups. However, we found that this spread our data too thin, as our dataset contained 18 unique airlines and 380 unique origin/destination codes, yielding a total of $18 \times 380 \times 380 = 2599200$ different groups, exceeding the number of data points we had by almost threefold. In the end, we decided it would still be quite useful to users to have a reliability measure per airline and airport (as opposed to airline and route, which yielded too many groups to analyze), which resulted in taking the average over only $18 \times 380 = 6840$ groups, yielding results like those shown in Figure 3. This way, when booking flights, customers have a very clear idea which flights are the most likely to be delayed so they can be avoided.

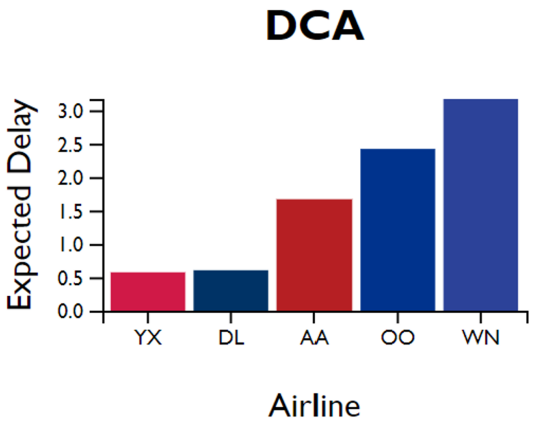

Figure 3: Expected flight delay by airline in DCA airport, as a reliability metric.

Conclusions & Discussion

Impacts and Limitations

We believe that this tool will be able to effectively communicate information to passengers and airlines, and help to ease stress and frustrations that are all too common in modern air travel, as it provides a clear, easy-to-use display, as shown in Figure 4 below.

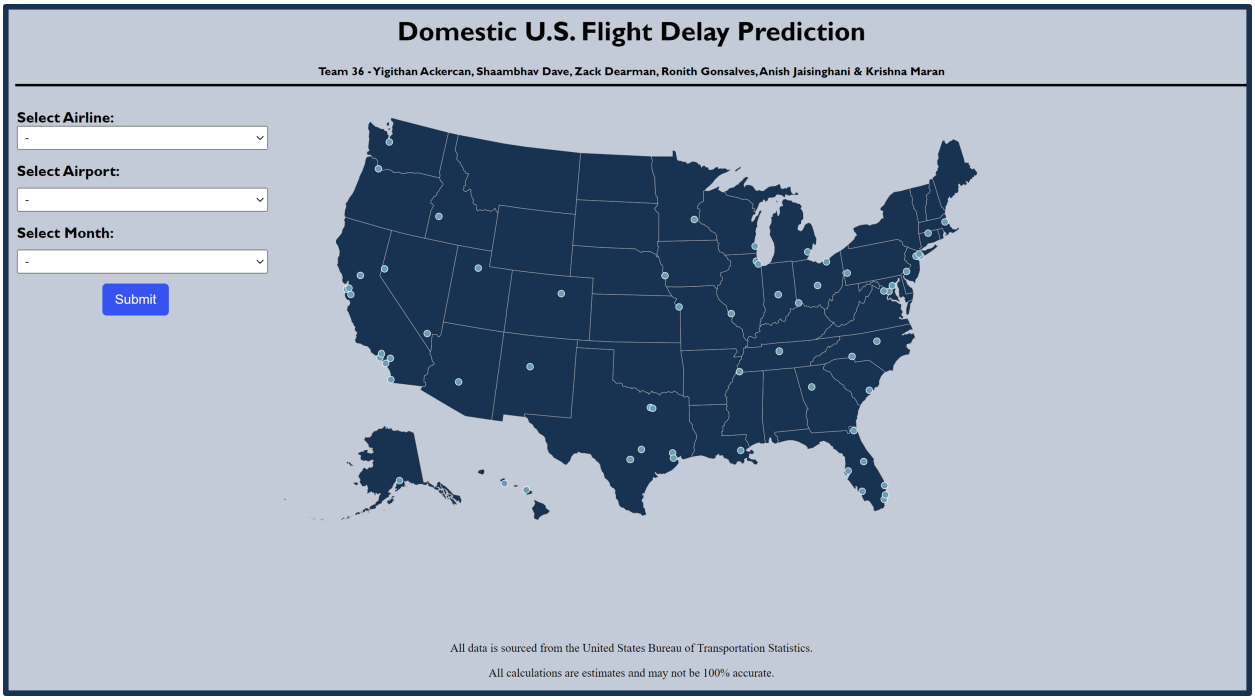

Figure 4: Interactive visual created for our project.

The accuracy of our is limited by the quality and scope of the data considered, as our model is only based on domestic flights from January 2019 through August 2023. By expanding the range of data, we would be able to improve the precision of anticipated delays, and limit the impact of outlying events, such as COVID-19.

Future Work

Future extensions to our project might include the consideration of other independent variables in their impact in flight delay, expansion to international flights leaving from the United States, and creating a feature to allow users to input their flight ID, and receive feedback specific to their flight. By considering additional factors, such as weather, arrival airport, or time of day, the model will be able to provide a prediction more relevant to a user's situation. Second, in considering international flights, we are able to expand the market for our model, as it provides insight for the international sector of airlines and global travelers. Lastly, the ability to input a flight ID and be given delay insight specific to our flight would be an interesting and desirable feature for many passengers, especially in the busy holiday seasons.

All team members have contributed a similar amount of effort to this project.

References

[1] M. K. Asfe, M. Jangizehi, M. N. Shahiki Tash, and N. M. Yaghoubi. Ranking different factors influencing flight delay. *Management Science Letters*, 4(7):1397–1400, 2014. doi: 10.5267/j.msl.2014.6.030.

[2] M. Ball, C. Barnhart, and M. Dresner. Total delay impact study : a comprehensive assessment of the costs and impacts of flight delay in the united states. *National Transportation Library*, 2010.

[3] L. Belcastro, F. Marozzo, D. Talia, and P. Trunfio. Using scalable data mining for predicting flight delays. *ACM Trans. Intell. Syst. Technol.*, 8(1), 2016. ISSN 2157-6904. doi: 10.1145/2888402.

[4] L. Carvalho, A. Sternberg, L. M. Gonçalves, A. B. Cruz, J. A. Soares, D. Brandão, D. Carvalho, and E. Ogasawara. On the relevance of data science for flight delay research: a systematic review. *Transport Reviews*, 41(4):499–528, 2021. doi: 10.1080/01441647.2020.1861123.

[5] M. Dai. A hybrid machine learning-based model for predicting flight delay through aviation big data. *Sci Rep*, 14:4603, 2024. doi: 10.1038/s41598-024-55217-z.

[6] Y. Ding. Predicting flight delay based on multiple linear regression. *IOP Conference Series: Earth and Environmental Science*, 2017. doi: 10.1088/1755-1315/81/1/012198.

[7] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou, and D. Zhao. Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, 69(1): 140–150, 2020. doi: 10.1109/TVT.2019.2954094.

[8] A. M. Kalliguddi and A. K. Leboilluec. Predictive modeling of aircraft flight delay. *Universal Journal of Management*, 2017. doi: 10.13189/ujm.2017.051003.

[9] Y. J. Kim, S. Choi, S. Briceno, and D. Mavris. A deep learning approach to flight delay prediction. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, 2016. doi: 10.1109/DASC.2016.7778092.

[10] M. Lambelho, M. Mitici, S. Pickup, and A. Marsden. Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *Journal of Air Transport Management*, 2019. doi: 10.1016/j.jairtraman.2019.101737.

[11] Q. Li and R. Jing. Flight delay prediction from spatial and temporal perspective. *Expert Systems with Applications*, 2022. doi: 10.1016/j.eswa.2022.117662.

[12] M. Mamdouh, M. Ezzat, and H. A. Hefny. A novel intelligent approach for flight delay prediction. *J Big Data*, 10:179, 2023. doi: 10.1186/s40537-023-00854-w.

[13] J. J. Rebollo and H. Balakrishnan. Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, 2014.

[14] Y. Tang. Airline flight delay prediction using machine learning models. Association for Computing Machinery, 2022.

[15] M. A. Vinayak Deshpande. The impact of airline flight schedules on flight delays. *Inform PubsOnline*, 2012. doi: 10.1109/TVT.2019.2954094.

[16] Y. Wang, M. Z. Li, K. Gopalakrishnan, and T. Liu. Timescales of delay propagation in airport networks. *Transportation Research Part E: Logistics and Transportation Review*, 161:102687, 2022. ISSN 1366-5545. doi: https://doi.org/10.1016/j.tre.2022.102687.

[17] C.-L. Wu and K. Law. Modelling the delay propagation effects of multiple resource connections in an airline network using a bayesian network model. *Transportation Research Part E: Logistics and Transportation Review*, 122:62–77, 2019. ISSN 1366-5545. doi: https://doi.org/10.1016/j.tre.2018.11.004.

[18] B. Yu, Z. Guo, S. Asian, H. Wang, and G. Chen. Flight delay prediction for commerical air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, 2019. doi: 10.1016/j.tre.2019.03.013.