PROJECT FINAL REPORT: VISUALIZING FLIGHT DELAYS

Team 154: Kaleb Boyd, Jacob Burdick, Jordan Gressel, Christopher Oyer

Introduction

Every year flight delays result in billions of dollars and countless hours lost for airlines, businesses, and passengers. Delays diminish the efficiency of the air traffic network and put strain on passengers, airline crew members, and other personnel. Much research has been conducted to better understand the causes of flight delays in an attempt to predict flight delay occurrence and better analyze delay propagation across the air traffic network. These studies have often included complex models that lack a robust visual component, making it difficult for most people to comprehend and draw meaningful conclusions from.

Our project's goal is to address this concern by creating a tool to visualize the factors that contribute to flight delays, showing how they are manifest over both time and geography. By incorporating a visual component, we aim to help individuals, businesses, and airlines better understand the factors driving flight delays and provide a foundation that can be built upon to create additional, more comprehensive tools that aid in analyzing and addressing the flight delay problem.

Literature Survey

Our team reviewed multiple studies related to the flight delay problem. A brief summary of our literature research is as follows:

Aljubairy describes the relatively new concept of utilizing real-time IOT data to predict flight delays. While this method is too time complex for us to implement, the paper did provide an example of a geographic visualization built from the data collected, which we can draw on when implementing our own visualization.

Brueckner focused on simplifying the analysis of flight delay propagation by focusing on the immediate cause of a delay. While this method does not dig deep to identify root causes, it is a quick way to get effective results. We can follow this pattern in developing our visualization to ensure it is easily understood and timely implemented.

Cai created a neural net to model the evolution of a flight network's graph structure over time. The model structure itself is complicated, however the combination of time series and graphs can help guide us as we create a temporal element for our own visualization.

Carvalho summarizes the methods described across several research papers to predict and analyze flight delays. The paper does not go in depth on the details of each method but gives our team a good foundation for how we can design our own visualization and ensure it is distinct from what is commonly done across the industry.

Chiraphadhanakul modeled delay propagation via an algorithm similar to a multi-dimensional Gantt chart to effectively show how delays affect slack across an air traffic network. As the model optimizes the worst-case scenario its predictions are too conservative, but combining this method with other models could lead to improvements. As we lack data regarding slack for airlines, we don't have specific plans to incorporate these ideas in our model.

Gopalakrishnan applies techniques used to detect outliers in graph signals for airport delay data. This is a great algorithm that, while currently narrow and poorly visualized, could be made much more interesting if visualized to show named cities, or plotted on a map. It has the potential to be easily implemented and used with other ideas we might have.

Li uses a Bayesian network to model how delays propagate and which airports typically delay other airports. While not very visual, the concept is useful by showing how delays are interconnected and could come in use as we attempt to visualize the geographical relationships between delays.

Mitsokapas uses probability density functions to analyze the impact of flight delays. This method is aligned with our goal to model the severity of flight delays and we will explore it further to see what can be applied to our visualization.

Peterson explores the economic costs of flight delays as well as the benefits of improving them. This article is excellent at showing why this problem is so important to so many groups of people and the kind of impact solving it can have on the nation's economy. We can use information found in this article to estimate the economic impact that our visualization can have.

Wang used an agent-based model to predict flight delays, with a focus on predicting time variables. The model is limited to one year and could be expanded to a larger data set to make it more effective. We could potentially use some of the ideas presented as we consider the time aspect of our visualization.

Wu explains the Delay Propagation Tree (DPT), a method for predicting delay propagation based on a single aircraft. While only applied to a limited number of flights, the paper demonstrates the effectiveness of solving a complex problem with a relatively simple approach. Ideas like this can help us reduce the complexity of our model if needed.

Zhou analyzed delay propagations using a causal analysis with key indicators, focusing on relationships between airports and correlations with time. The paper provides a framework with which we can pattern our own analysis as we couple it with a visual component.

Proposed Method

Rather than focusing on simply predicting if a flight delay will or will not occur, our method focuses on calculating a probability that a flight delay will occur given a set of factors. By focusing on probabilistic responses as opposed to a binary response, we gain a better understanding of the severity as well as the occurrence of flight delays for a particular airline, airport, and time of year. Such a response coupled with a visualization allows for better comparison of various factors contributing to flight delays and can reveal more nuanced patterns and correlations between these factors.

Probability Model

Our first step was to create a model to determine the probability of a flight delay. After performing exploratory data analysis and running small scale tests for a variety of models, our team settled on creating a simple random forest classification model focused on using the airport of origin and other features to predict the probability of a flight delay.

Based on our exploratory data analysis we used Python to filter and clean the data in preparation for building the model. We considered any delays greater than 1,000 minutes as outliers and removed these from the data. We next created a binary variable to indicate if a flight was delayed or not, assigning the value of 1 to any delay greater than five minutes, and assigning a value of 0 to all others. We removed all unnecessary factors from the data leaving us with the remaining factors we planned to incorporate into our model as predicting variables. These factors are month (numerical), airport origin (categorical), and airline (categorical). The airport origin and airline variables were one hot encoded to dummy variables so they could be used in the model. One million records were randomly sampled to be treated as the training dataset.

Once the data was prepared, we used Python to fit a random forest classification model to the dataset using hyper-parameter optimization through the grid search method. Different combinations of the the "n_estimators" (number of trees in the forest), "max_depth" (number of splits that each decision tree can make) and "min_samples_leaf" (minimum number of samples required at each leaf node) parameters were tested in increments of 5 between 5 and 30 to find the best accuracy score through the cross validation method. After testing each combination, the optimal parameters identified were "max_depth" =20, "min_samples_leaf" =5 and "n_estimators" =10 with a model classification accuracy score of 75.39% . The random forest classifier model was fitted with these parameters and then output to a pickle file to be utilized in the visualization. To generate predictions using the model we used the "predict_proba" method found in the scikit-learn Python package, which predicts the probability of a delay (class 0 or 1) based on values for our selected predictors.

Map Visualization

To create our visualization we used a combination of the Dash and Plotly packages in Python. The visualization itself is a graph of the air traffic network in the United States overlaid on a map of the world (in order to capture some of the farther reaching flights to places like the islands of the Pacific). Each airport that was found in the dataset was a node while each unique route (flight from an origin to destination airport) was an edge of the graph.

The interface includes two drop down menus that are used to set the values of the airline and month of year parameters used in the probability model to calculate the likelihood of a delay. The user also has the ability to select one of the nodes (i.e. airports) of the graph which sets the value of the final parameter for the probability model and executes it. The output of the model is displayed directly below the drop down menus and shows the probability of a flight being on time as well as the probability of a flight being delayed more than five minutes for the criteria specified.

In addition, when the user selects an airline, all airports that the airline does not fly into or out of are hidden on the visualization, leaving only the airports the airline travels to showing. All flight paths of which the selected airport serves as the origin are also kept in the visualization while any other flight paths for other airports are hidden. The user also has the ability to click and drag the map and to zoom in or out to better view the visible portions of the network. The user can hover over any visible airport to see its name.

The underlying data, visualization, and probability model are all run locally on a user's computer via a python virtual environment.

Experiments and Evaluation

When we started work on the project, the dataset we initially chose only covered a one-year period and had a small number of features to choose from. We decided to look for a more robust data set and eventually found one that included five-years' worth of flight data (a total of over 29 million flights) along with 61 distinct features captured for each flight. The data itself was very clean and was tested/validated by each team member during the exploratory data analysis phase of our project. Due to the large size of the data files, we used parquet files in order to store and read the data.

After creating the random forest classification model and fitting it with the optimum parameters, we evaluated its performance using the Receiver Operating Characteristic (ROC) curve. We used a test dataset consisting of 10,000 flights randomly sampled from the dataset and excluded from the training data set. The scikit-learn "predict_proba" method was utilized to distinguish how accurate our model was at predicting a classification (0 or 1) for a flight compared to the true classifications for the flights in the test data set. Our model had an Area Under the Curve (AUC) score of 0.61 which indicates that the model performs poorly when trying to determine the probability of a flight cancellation. This is a slight increase over the AUC score of 0.6 for a model without hypertuning the parameters, but still less than a target score of 0.7 to 0.8 which would indicate an acceptable level of performance for the model. To improve the accuracy of future iterations of our model, we could test different features such as the destination airport, flight path, time of day, or weather. Creating multiple models based on a specific key feature with additional supporting features (e.g. origin airport as a key feature with different combinations of supporting features) could also help improve the model's prediction accuracy.

To test the visualization, each team member ran it on their own computer and tested various components of the user interface (basic functionality, ensuring airports and flight paths showed up correctly, ensuring zoom/drag functionality was present, etc.). While doing so we uncovered a number of details that needed to be adjusted including:

- Increasing the overall speed of loading the visualization.
- Fixing an auto-recenter issue when a user clicked on an airport. The map would jump back to a predetermined center point rather than stay focused on where the user clicked.
- Updating the map to hide airports and flight paths that were not utilized by the selected airline.
- Updating the size of the nodes (airports) to make them easier to spot by the user.
- Adjusting the CSS styling to make the page dynamic to resizing

In testing the visualization we observed the following characteristics regarding flight delays:

- The chance of delay was fairly consistent for an airport regardless of the origin (with minor exceptions).
- The visualization does not provide a simple way for comparing flight delays across airlines.

Results and Conclusions

Overall, the model and visualization we created provide a solid foundation upon which we or others could build and refine to provide additional insights regarding flight delays and help identify solutions to reduce this problem across the air traffic network.

The probability model is a big element that would need improvement for future iterations. Ideas for improving the accuracy of the model predictions include:

- Implementing some of the ideas already suggested in the *Experiments and Evaluation* section.
- Narrowing the scope of the model to a subset of airlines or airports.
- Using variable selection techniques (such as lasso or elastic net regression) to identify the features that have the most influence on flight delays.
- Revisiting models we considered earlier on in our project to see if they would yield better results.

The visualization element of our project has a lot of potential to be improved and enhanced. Ideas for improvement include:

- Including more/relevant drop downs to input features into the predictive model.
- Include a call-out box that displays statistics and other relevant information for a selected airport.
- Adding the ability to select a flight path and view information regarding that specific path.
- Adding a heat map to show how the probability of delay changes with geography.
- Incorporating a way to show how delays propagate through the air traffic network.
- Creating a similar visualization for flight cancellations and deviations.
- Adding dynamic charts, in addition to the map, that display time-series statistics for the chosen airport or airline.

Aside from these two elements, our team also worked on a PyTorch model to implement a time series graph convolution model. The data was a time series of adjacency matrices: each matrix was filled with delay times between airports within a time window. The added complexity of the data structure and model meant this model was not fully ready within the project timeline, but finishing this model would add a powerful way of using the intrinsic structure of the data to improve the predictions.

All team members contributed a similar amount of effort while working on this project.

WORKS CITED

Aljubairy, A., Shemshadi , A., & Sheng, Q. Z. (2017). Real-Time Investigation of Flight Delays Based on the Internet of Things Data. In *Advanced Data Mining and Applications: 12th International Conference, ADMA 2016* (Vol. 10086, pp. 788–800). essay, SPRINGER.

Brueckner, J. K., Czerny, A. I., & Gaggero, A. A. (2021, July). *Airline delay propagation: A simple method for measuring its extent and determinants*. CESifo. Retrieved February 27, 2023, from https://www.cesifo.org/en/publications/2021/working-paper/airline-delay-propagation-simple-method-measuring-its-extent-and

Cai, K., Li, Y., Fang, Y.-P., & Zhu, Y. (2022). A deep learning approach for flight delay prediction through time-evolving graphs. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 11397–11407. https://doi.org/10.1109/tits.2021.3103502

Carvalho, L., Sternberg, A., Maia Gonçalves, L., Beatriz Cruz, A., Soares, J. A., Brandão, D., Carvalho, D., & Ogasawara, E. (2020). On the relevance of data science for Flight Delay Research: A systematic review. *Transport Reviews*, *41*(4), 499–528. https://doi.org/10.1080/01441647.2020.1861123

Chiraphadhanakul, V., & Barnhart, C. (2013). Robust flight schedules through Slack re-allocation. *EURO Journal on Transportation and Logistics*, 2(4), 277–306. https://doi.org/10.1007/s13676-013-0028-y

Gopalakrishnan, K., Li, M. Z., & Balakrishnan, H. (2019). Identification of outliers in graph signals. *2019 IEEE 58th Conference on Decision and Control (CDC)*, 4769–4776. https://doi.org/10.1109/cdc40024.2019.9029478

Li, Q., & Jing, R. (2021, May 19). Characterization of delay propagation in the Air Traffic Network. *Journal of Air Transport Management.* Retrieved February 28, 2023, from https://www.sciencedirect.com/science/article/pii/S0969699721000582

Mitsokapas, E., Schäfer, B., Harris, R., & Beck, C. (2020). Statistical characterization of airplane delays. *Nature Portfolio*, *11*. https://doi.org/10.21203/rs.3.rs-133115/v1

Peterson, E. B., Neels, K., Barczi, N., & Graham, T. (2013). The Economic Cost of Airline Flight Delay. *Journal of Transport Economics and Policy* , 47(1), 107–121. https://www.jstor.org/stable/24396355

Qiang Li, Ranzhe Jing (2021) Characterization of delay propagation in the air traffic network *Journal of Air Transport Management* Volume 94, July 2021, 102075 https://www.sciencedirect.com/science/article/pii/S0969699721000582

Wang, C., Hu, M., Yang, L., & Zhao, Z. (2021). Prediction of air traffic delays: An agent-based model introducing refined parameter estimation methods. *PLOS ONE*, *16*(4). https://doi.org/10.1371/journal.pone.0249754

Wu, W., & Wu, C.-L. (2018). Enhanced delay propagation tree model with Bayesian network for modeling flight delay propagation. *Transportation Planning and Technology*, *41*(3), 319–335. https://doi.org/10.1080/03081060.2018.1435453

Zhou, F., Jiang, G., Lu, Z., & Wang, Q. (2022). Evaluation and analysis of the impact of airport delays. *Scientific Programming*, *2022*, 1–8. https://doi.org/10.1155/2022/7102267