Predicting Flight Delays Using Neural Networks

Garrett Tate, April 2021

Delayed flights present a significant challenge to travelers and airlines alike, with 19% of US domestic flights delayed in 2019[1] at an estimated total cost of $33 billion that year[2]. Airlines incurred a cost of approximately $8.3 billion, with causes ranging from increased crew costs during unexpected reschedules to reduced aircraft utilization from buffers built into flight scheduling. Passengers meanwhile collectively incurred costs totaling a whopping $18.1 billion, with causes varying from actual missed connections to the cost of travelling earlier than necessary in anticipation of delays. Meanwhile, approaches to plan for flight delays remain rudimentary and inefficient. Airlines use blanket minimum connection times for all flights in a given airport; customers often employ static rules of thumb for the length of every layover they schedule; and considerations for the probability of delays in a given airport or on a given airline are typically left to the anecdotal experiences of frequent fliers. Therefore, there exists significant opportunity to improve the efficiency of airline travel by estimating the probabilities of delays for individual flights, thereby allowing airlines to schedule more efficiently, third-party bookers to better serve customers with more reliable itineraries, and consumers to plan critical business trips and precious vacations more confidently.

Here I have constructed a machine learning modeling procedure that predicts the probability of flight delays 1-2 months in advance. The full Python code for modeling and evaluation can be found at github.com/Tate-G/portfolio. The model only uses information about the flight schedule, the airline, and the origin and destination, all of which is known at the time of booking. Data for training and testing these models are taken from the Reporting Carrier On-Time Performance database from the US Bureau of Transportation Statistics[3], a monthly database with detailed characteristics and delay information for every domestic US flight. After data import I clean and reformat necessary features, including one-hot encoding categorical variables (such as airlines and airports), representing cyclicity in repeating time variables using trigonometric functions (such as with the time of day), and normalizing continuous variables (such as scheduled flight duration). I train a neural network model using Tensorflow and Keras to classify whether flights are delayed. I consider a flight delayed if the arrival time is fifteen minutes late or more, following the convention used by the Federal Aviation Administration. Memory constraints on my own hardware required that I limit the size of the training data, so I restrict my modeling to flights to and from the one hundred busiest airports and I randomly select a subset of training flights over several years. Each neural network model is trained on about four million flights during twenty-four consecutive months, validated during training on all flights from the following month (about half a million), and finally tested on all flights during the month following the validation period. I repeat this procedure with a rolling window of training, validation, and test months to estimate the potential variability in model performance if it were deployed. I compare each model's performance against a baseline model in which flight delays are predicted randomly in proportion to the total flight delays in the training dataset.

My own expectation before modeling was that exceptional levels of precision predicting flight delays may not be possible, considering that airlines are actively reviewing performance to try to prevent delays and many delays are caused by unpredictable events including the weather. Nevertheless, I also expected that given typical patterns in seasonal climate, common cycles in the busiest travel times, and varying practices for each airline or airport, models would provide a notable uplift compared to randomly guessing delays. Given the scale of the economic impact of US flight delays, even improving delay predictions by a few percent could save airlines and passengers billions of dollars.

---

[1] United States Bureau of Transportation Statistics, https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E

[2] United States Federal Aviation Administration, Cost of Delay Estimates 2019, https://www.faa.gov/data_research/aviation_data_statistics/media/cost_delay_estimates.pdf

[3] Data available at https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time

Testing against 2019 data shows that my neural network modeling procedure provides significant improvement for predicting individual flight delays compared to the baseline of assuming a constant delay probability. Figure 1 shows results from tests on flights in 2019, with ROC area under the curve (AUC) ranging from 0.62 – 0.69 compared to a baseline AUC of 0.50. Model AUC was highest from May to August 2019 and AUC was lowest in February and November.

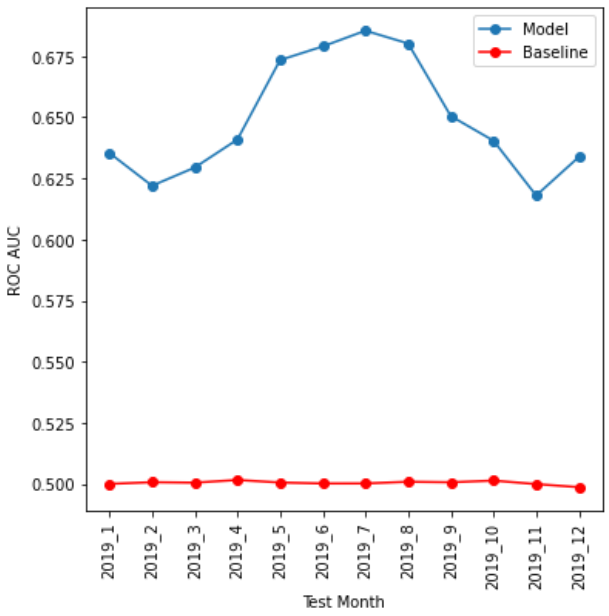

Figure 1: ROC Area Under the Curve (AUC) for neural network models and baseline models tested on each month in 2019.

If this model were deployed, developers would have wide latitude to favor either precision or recall as best suits the application. For instance, airlines might favor a model that maximizes precision by identifying specific flights that are very likely to be delayed, since it may be very expensive for an airline to reschedule flights within a complex and interconnected system. Customers on the other hand may be most concerned about the cost of missing a connection or an event at their destination, therefore customers may prefer models maximizing recall that identify as many reasonably likely delays as possible. While testing the model on 2019 flights, a model threshold favoring precision correctly identified an average of 338 delayed flights a month (with an average precision of 0.52 and an average recall of 0.003). Testing the same models with thresholds that favor recall for 2019 correctly identified an average of 61,819 delayed flights a month (with an average precision of 0.27 and an average recall of 0.62).

This modeling procedure encountered more muted success when tested against flight data in 2020 during the COVID-19 pandemic. After yielding an AUC of 0.58 – 0.60 in January to March 2020, the AUC in April plummeted to 0.46 (that is, worse than random guessing). The AUC from May through December 2020 ranged from 0.50 to 0.61. Although notably worse than the performance in 2019, this modeling still provided predictions at least as good and usually better than the baseline during almost all of 2020. Given the fundamental changes in airline travel during the pandemic (for instance total flights dropped 36% from 2019 to 2020[4]), and given the potential economic benefits of even marginal improvements in flight delay forecasting, even the smaller gains offered by this modeling during the pandemic are providing a useful result.

To identify which features are most important toward predicting flight delays, I calculate the permutation feature importance[5,6] for each feature in the model. For one feature at a time, I randomly re-order the values for that feature while leaving all other features unchanged, then compare model performance with the original data versus performance with the permuted data that has one randomly re-ordered feature. Randomly permuting the values of very important features should noticeably

[4] For US reporting carriers, according to the US Department of Transportation February 2021 Air Travel Consumer Report, https://www.transportation.gov/individuals/aviation-consumer-protection/february-2021-air-travel-consumer-report

[5] Fischer, Rudin, and Dominici, 2019, All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously, Journal of Machine Learning Research, https://arxiv.org/abs/1801.01489v5

[6] Molnar, 2021, Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/

decrease the AUC of the predictions, as randomization will break the relationships that the model depends upon to predict delays. Permutation feature importance values are calculated as:

$$Feature\ Importance = \frac{1 - AUC_{feature\_randomized}}{1 - AUC_{original\_data}}$$

The most significant factors revealed by permutation feature importance in one example model are shown in Figure 2. Flight departure and arrival times are the most individually significant predictors of delays. Figure 3 shows how the predicted delay probability varies through the day, with lowest probability in the morning and highest probability in the evening. Among airlines, Delta, United, and Frontier are the three most important to the model. Figure 4 compares the distributions of delay probabilities for Delta and Frontier. Only 24% of Delta flights have a delay probability higher than the median for all flights, whereas 93% of Frontier flights have a delay probability higher than the overall median. Among airports, San Francisco International is the most important to the model, and 84% of flights arriving there have a delay probability above the overall median. Meanwhile, only 13% of flights originating from Hawaii have a delay probability above the median for all flights. Although permutation feature importance does not capture the interconnected relationships between factors in the model or the full predictive power of the model, highlighting feature importance during model deployment would provide useful new information for customers considering flight options and for airlines searching for areas of their business that are ripe for process improvements.

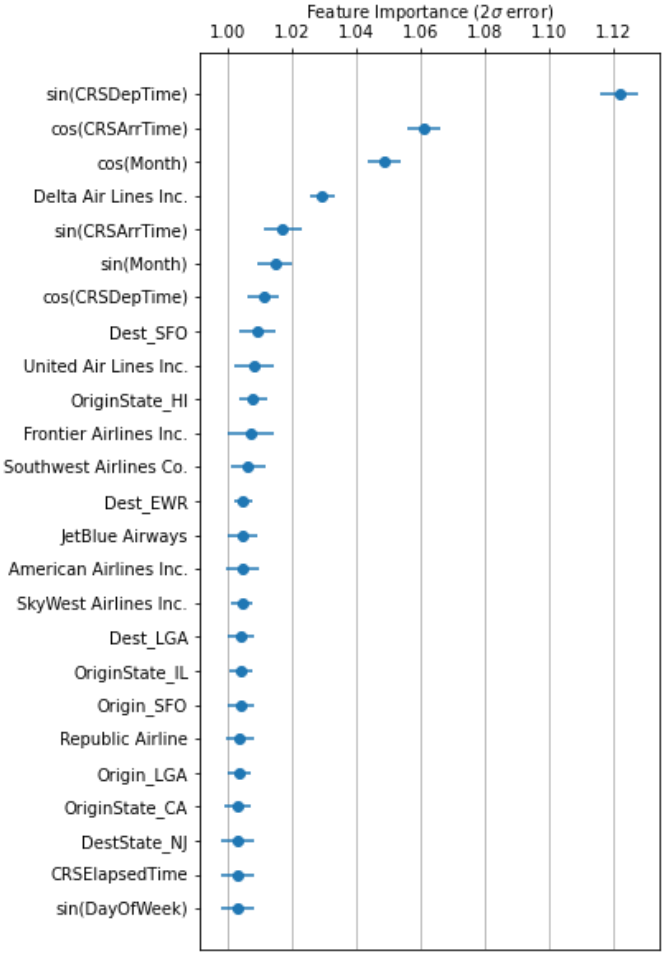

Figure 2: Feature importance for the 25 most important features as calculated for the model tested on December 2019. A feature importance of 1 means that randomizing feature had no detectable effect on model performance. Note that cyclical time features (time of day, month, etc.) have been transformed using trigonometric functions to represent their cyclicity.

Opportunities abound for impactful real-world deployment of flight delay models such as this. A variation of this modeling procedure focusing on flights from a particular airline would allow one to pinpoint specific flights most likely to be delayed months in advance, allowing rescheduling of select high-risk flights before many passengers even book their travel. Or, booking services could differentiate their product by showing customers which flights

are most likely to be delayed, increasing customer confidence when purchasing travel. Another variation of this modeling could integrate delay probabilities with scheduled layover times when booking connecting flights, reducing the number of missed flight connections. Some potential challenges remain before deployment, such as effective communication of delay probabilities. Many people often underestimate the likelihood of events that have low probability[7,8], so effectively communicating the odds of a delay will be an important aspect of serving customers. Among other potential challenges is the uncertainty of model accuracy during the remainder of the COVID-19 pandemic and after the pandemic is over. Testing on 2020 data reveals reduced model performance compared to 2019, therefore reported delay probabilities may not be as accurate during the rest of the pandemic either. Additionally, after the pandemic is over, models should be tested further to determine if their performance is greater with or without including training data from the height of the pandemic. Nevertheless, as we exit the COVID-19 pandemic there is a prime opportunity to begin incorporating more robust delay predictions. At this moment airlines are setting new flight schedules and customers are emerging from isolation eager to travel, so now is the time to create more efficient flight schedules, capture customer loyalty, and work toward saving billions of dollars for airlines and passengers in the United States.

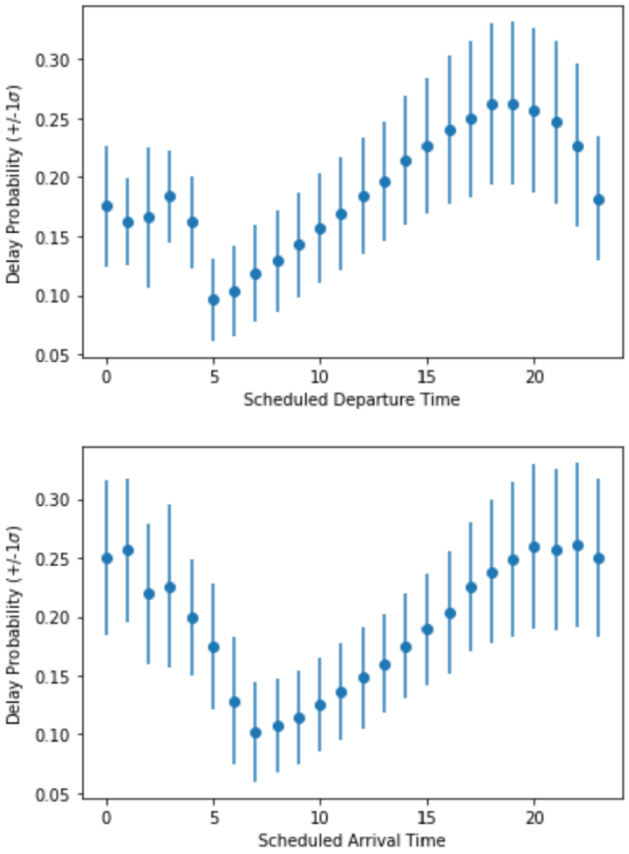

Figure 3: Modeled delay probability by departure and arrival time for flights in the training data of the December 2019 model.

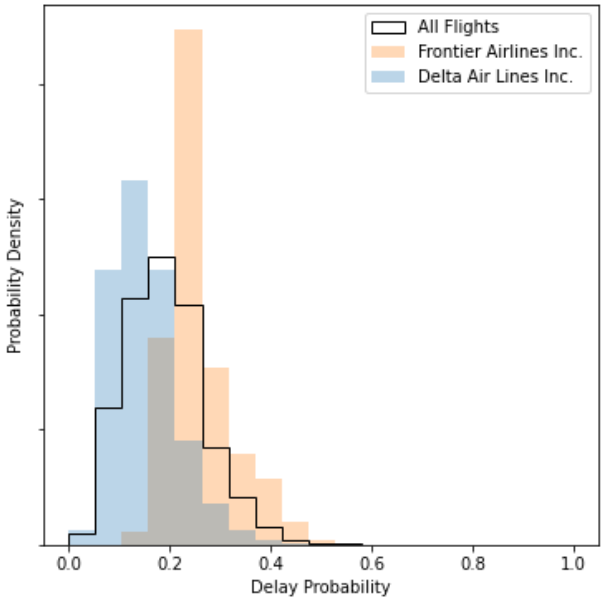

Figure 4: Modeled flight delay probability distributions for all flights and for select airlines in the training data for the December 2019 model.

[7] Caponeccia 2006, Strategies for Effective Communication of Probabilities, https://cebra.unimelb.edu.au/__data/assets/pdf_file/0012/2221203/0608.pdf

[8] Fischhoff, Brewer, and Downs, 2011, Communicating Risks and Benefits: An Evidence-Based User's Guide, United States Food and Drug Administration, https://www.fda.gov/files/about%20fda/published/Communicating-Risk-and-Benefits---An-Evidence-Based-User%27s-Guide-%28Printer-Friendly%29.pdf