Proceedings of Machine Learning Research 82 (2017) 45-51, 4th International Conference on Predictive Applications and APIs

Prediction and Uncertainty Quantification of Daily Airport Flight Delays

Thomas Vandal TJ.VANDAL@GETFREEBIRD.COM
Max Livingston MAX@GETFREEBIRD.COM
Camen Piho CAMEN@GETFREEBIRD.COM
Sam Zimmerman SAM@GETFREEBIRD.COM

Abstract

One in four commercial airline flights is delayed, inconveniencing travelers and causing large financial losses for carriers. The ability to accurately predict delays would make travelers' lives easier and save airlines money. In this work, we approach the problem of predicting flight delays using a Variational Long Short-Term Memory (LSTM) model. The model is trained to predict aggregate daily delays for U.S. airports using a combination of continuous and discrete variables, including weather, airport characteristics, and congestion. Monte Carlo Dropout, a Bayesian Deep Learning technique based on variational inference, is incorporated to provide planners with a well-calibrated prediction interval. We show that our Variational LSTM results in an average median absolute error of 5.8 minutes per day across 123 airports in the United States. Moreover, results show that predictive uncertainty is well explained through a calibration analysis.

1. Introduction

Commercial flight delays in the U.S. airspace are a pervasive and expensive problem. A large-scale study by the U.S. military found approximately 25% of flights were delayed in the late 2000s (Fleming, 2009). Recent analysis done by a large travel technology company estimates the economic impact of these disruptions to be $60B and growing (Gershkof, 2016), representing over 8% of global airline revenue. With ridership in the U.S. expected to double or triple in the next 15 years and few definitive plans for infrastructure improvements, this economic impact will likely increase.

In this paper, we focus on our work predicting aggregate per-airport daily delays. A forecast of the average delay intensity at an airport for a given day can help travelers, air traffic controllers, airport operations, and travel agents stay informed about airport operations and make proactive decisions (Demuth et al., 2011). This strategy is also consistent with the literature, which typically analyzes aggregate delays (Gopalakrishnan and Balakrishnan, 2017).

There are two pieces of existing research in this domain that are similar to ours. The first, (Gopalakrishnan and Balakrishnan, 2017), compares performance of two different neural network techniques (a Multi-Layer Perceptron (MLP) and a General Regression Neural Net (GRNN)), traditional machine learning techniques, and a domain-specific technique for the problem of predicting aggregate origin-destination level and origin level delays.

Using a dataset of the top 30 airports (the FAA Core 30 list), they train the model to predict average outbound delays at an airport two hours ahead, given outbound delays now, outbound delays in the previous two hours, and the time of the day. By training an MLP with those features, they achieve a median absolute error of around 7 minutes. The second, Kim et al. (2016), uses a LSTM-RNN on daily flight delays. They focus on classification, setting a threshold for the average daily delay for an airport and training the RNN to predict whether a day's average delay would be above that threshold. They achieve an average classification accuracy of between 80 to 90%.

Our primary advancement over previous work is the modeling technique used, a Variational LSTM. This is a key feature of our approach and is crucial for Freebird's applied business aims. Because Freebird is interested in using these models in a risk management setting, we require a model that provides robust uncertainty metrics, not just point estimates. This requirement has limited the application of deep learning to the problem thus far. We leverage recent advances from (Gal and Ghahramani, 2016; Gal, 2016; Gal et al., 2017), using Monte Carlo dropout to obtain parameter variance estimates. To our knowledge, we are the first to apply Bayesian deep learning to the problem of predicting and quantifying flight delays in the U.S. airspace and the first to apply a Variational LSTM to any industrial problem.

2. Data Description

Predicting average daily flight delays for a given airport is a challenging task with many factors in play. However, many of these factors, such as operational malfunctions, are unpredictable in advance, which limits our feature set to airport level characteristics and weather variables. The features we use are as follows:

Continuous Features: Max Wind Gust, Avg. Wind Speed, Avg. Heat Index, Total Precipitation, Avg. Pressure, Total Snow, Avg. Temperature, Avg Visibility, # of Arrival Flights, # of Departing Flights, Previous Day's Delay
Categorical Features: Month, Day of Week, and Airport

The weather features were extracted from historical data provided by The Weather Company. The data we were provided included hourly statistics such as temperature, wind, snow, pressure, and many others. Variables were then aggregated from hourly to daily by either averaging, computing the max, or taking the sum. These features are then normalized across all airports before training.

Daily average flight delays and number of departing and arrival flight were extracted from a large proprietary data feed containing flight statuses for essentially all flights in the U.S. The number of arriving and departing flights were normalized per airport in order to preserve well distributed features. We note that these features are available approximately 200 days in advance given flight schedules and weather forecasts. We use data from July 1 2012 to July 31 2016, with a train/test split of 80%/20% . The final dataset consists of 123 airports over 4 years of daily samples, totalling 168,387 samples for training.

3. Method

In this section we describe the Bayesian LSTM architecture we developed to predict daily average flight delays per airport using both continuous and categorical features. Including an airport indicator variable allows us to leverage similar delay effects between airports, similar to a multi-task model. The model consists of 3-Layer LSTM network where data from the day of departure and the four previous days is used to predict average delay for the day of departure. First, we embed the categorical variables from 7 days to 3 features, 12 months to 3 features, and 123 airports to 5 features. The weights for embedding are learned during training. Using a grid search with a range of dimensions, we found that these embedding dimensions provided a good trade-off between complexity and over-fitting. Categorical variables embedded to dimensions 3, 3, and 5 are concatenated with the 11 continuous features, including weather and airport congestion, resulting in 22 total features that are fed into the LSTM.

Following (Gal and Ghahramani, 2016; Gal, 2016), we assume $\mathbf{y} \sim \mathcal{N}\big(\mu(\mathbf{x}), \sigma^2(\mathbf{x})\big)$ such that $[\mu(\mathbf{x}), \sigma(\mathbf{x})] = f^{\boldsymbol{\omega}}(\mathbf{x})$ where f is an LSTM network with two hidden layers of 128 units each. Approximate variational inference is applied over all the weights, $\boldsymbol{\omega}$, using Monte Carlo Dropout. As is crucial in Variational LSTMs, the same dropout mask is used at each step and for all weights (rather than dropping different weights at each time step). The corresponding negative log-likelihood is then written as:

$$\begin{aligned}\mathcal{L}(\theta) &= -\sum_{i\in\mathcal{S}} \log p\big(\mathbf{y}_i | f^{\hat{\boldsymbol{\omega}}}(x)\big) + \mathrm{KL}(q(\boldsymbol{\omega})||p(\boldsymbol{\omega})) \\ &= \frac{1}{D}\sum_{i\in\mathcal{S}} \frac{1}{2}\exp(-s_i)^{-1}||\mathbf{y}_i - \mu(\mathbf{x}_i)|| + \frac{1}{2}s_i + \frac{1-p}{2N}||\boldsymbol{\omega}||_2\end{aligned} \tag{1}$$

with $s_i = \log(\sigma^2)$ to enforce a positive variance where $\mathcal{S}$ consists of a sample batch of size D, θ denoting all parameters, and the dropout probability p. Using this approach, the first two moments of the predictive distribution have the following unbiased estimates:

$$\begin{aligned}E[\mathbf{y}] &\approx \frac{1}{K}\sum_{k=1}^{K} \mu^{\hat{\boldsymbol{\omega}}_k}(\mathbf{x}) \\ Var[\mathbf{y}] &\approx \frac{1}{K}\sum_{k=1}^{K} \mu^{\hat{\boldsymbol{\omega}}_k}(\mathbf{x})^2 + \frac{1}{K}\sum_{k=1}^{K} \sigma^{\hat{\boldsymbol{\omega}}_k}(\mathbf{x})^2 - \Big(\frac{1}{K}\sum_{k=1}^{K} \mu^{\hat{\boldsymbol{\omega}}_k}(\mathbf{x})\Big)^2\end{aligned} \tag{2}$$

where $\hat{\boldsymbol{\omega}}_k$ refers to weight realizations of $\boldsymbol{\omega}$ after applying dropout while K denotes the number of Monte Carlo samples (we set $K = 30$ in our experiments). A constant dropout rate of 0.25 is used in all layers. We refer readers to section 4 in Gal and Ghahramani (2016) for a more detailed analysis of dropout and uncertainty quantification in recurrent neural networks.

4. Results

To test our Variational LSTM, as described above, we study its ability to produce accurate daily average delay predictions per airport as well as corresponding predictive uncertainty.

	MAE				RMSE			
	1	3	5	7	1	3	5	7
Airport								
ATL	3.60	4.07	4.54	4.67	7.90	8.11	8.32	8.37
DFW	3.88	4.66	5.16	5.82	11.13	11.44	11.61	11.82
JFK	4.36	5.31	5.27	5.40	13.85	13.85	13.82	13.89
LAX	2.95	3.44	3.58	3.77	4.59	4.92	5.07	5.25
ORD	6.20	7.63	8.35	9.14	13.03	13.93	14.47	15.01
All	5.84	6.60	7.22	7.71	10.64	11.41	12.03	12.56

Table 1: Predictive ability at major airports for 1, 3, 5, and 7-days ahead measured in minutes per day by Median Absolute Error (MAE) and Root Mean Square Error (RMSE).

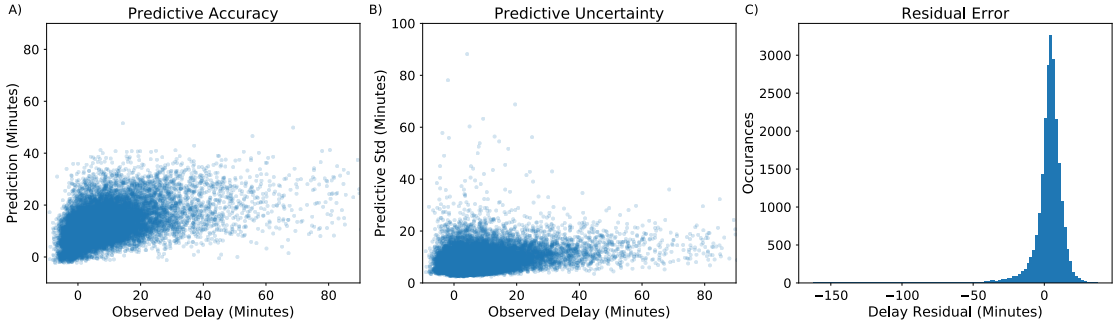

Figure 1: These results include test data only. (a) Observed average daily delay on the x-axis, which is truncated to 90 minutes) and predicted on the y-axis for 1 step ahead. (b) Standard deviation of the prediction by the observed delay (also truncated to 90 minutes). (c) Histogram of model residuals (Predicted - Observed)

All analyses are performed on a held-out test set (the last 20% of observations). To understand overall predictability, we visually compare the predictions and observations for each airport-date in Figure 1(a) and more quantitatively in Table 1. The long-tail distribution of delays results in a skewed error distribution, shown in Figure 1(a). In general, the model is not able to predict rare extreme delay events. For instance, the third largest average daily delay in our test set, 136 minutes, was on July 4th 2016 at John F. Kennedy (JFK) Airport in New York. These delays were caused by ISIS threats to three major international airports, including JFK. While this is just a single example, similar events happen, however infrequently, which are unreasonable to capture in the predictive means (as we show later on, we do reasonably well at accounting for the uncertainty via the predicted variance). If we remove observations with daily delay averages of over an hour, our RMSE for the test set drops from 10.64 to 9.21. To help understand the sensitivity to outliers, we report the Median Absolute Error (MAE) alongside the root mean squared error (RMSE) in Table 1 for predictions 1, 3, 5, and 7 days ahead for the top five airports by passenger count, as well as the validation dataset as a whole.

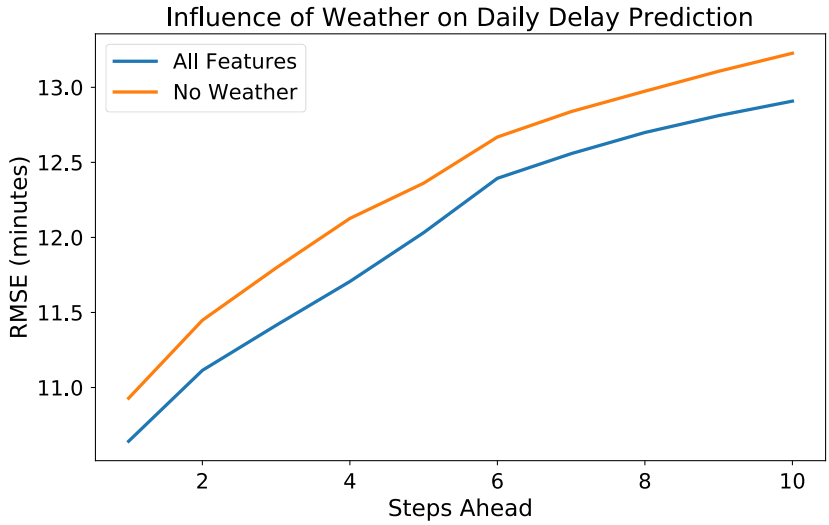

Figure 2: Comparison between model with (blue) and without weather features (orange) for 1 to 10 days ahead.

To gain insights into the importance of weather on delay prediction, we experiment with and without weather features. The model with weather outperforms the model without, with a decrease in RMSE of about 1 minute, shown in Figure 2, with roughly consistent improvement across prediction windows.

Lastly, we study the width and quality of predictive uncertainty. Figure 3 presents the calibration quality by measuring the frequency of observations within a given probability interval. For example, we expect 20% of the observations to fall between the estimated 40^{th} and 60^{th} percentiles. Too many observations outside of that percentile range would result in the curve being above the 45-degree line at 0.2. We clearly see that the uncertainty calibration is good at nearly all the airports. In contrast to the predictive mean, Figure 1 shows that predictive uncertainty measured by standard deviation does not increase dramatically with the observed delay. This all results in well-calibrated uncertainty with reasonable predictability.

5. Conclusion

This paper presented a novel Bayesian RNN-LSTM to predict aggregate flight delays in the US by airport. In addition, we also extended these deep learning models to extract stable uncertainty quantification for each prediction using Monte-Carlo Dropout techniques. Our analysis of the model showed that even at the daily aggregate level, flight delays are quite challenging to predict, with an average RMSE above 10 minutes. This furthermore motivates the need for uncertainty quantification. The approximate predictive posterior of our model was shown to be strong with little error in calibration.

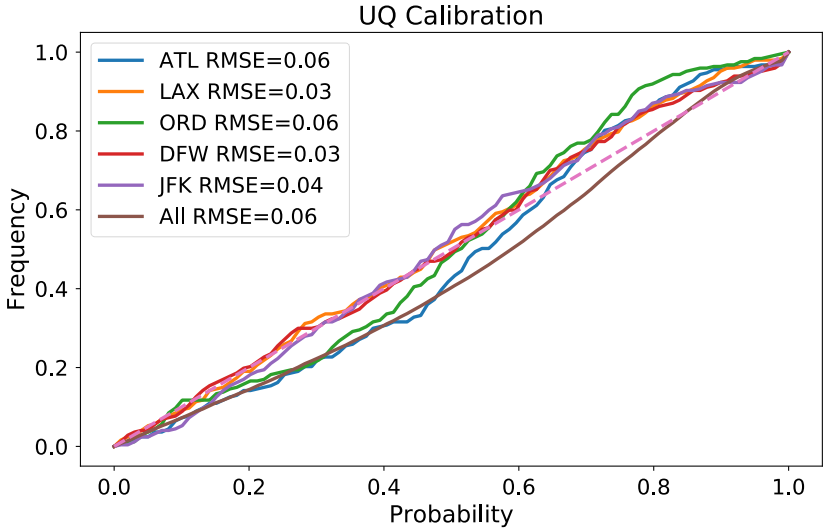

Figure 3: Uncertainty calibration measured by the frequency (y-axis) of observations within a given predictive interval (x-axis). The ideal calibration is when $y = x$, represented by the dashed line. RMSE is the error between calibration and ideal.

Future work could improve the training process to allow for more granular predictions, such as hourly forecasts or even flight-by-flight forecasts informed by the daily predictive distribution. Augmenting the RNN with additional features to capture network structure would allow us to model phenomena such as cascading delays across connected airports, represented using properties of the network. Furthermore, incorporating more domain informative features such as the Homeland Security Advisory System threat level and natural language processing feeds of relative news and social media platforms. An additional avenue for future work is to extend this model to better characterize different types of flight delays using mixed-density models, and to meaningfully distinguish between different types of uncertainty (Kendall and Gal, 2017).

References

Julie L. Demuth, Jeffrey K. Lazo, and Rebecca E. Morss. Exploring variations in people's sources, uses, and perceptions of weather forecasts. *Weather, Climate, and Society*, 3(3): 177–192, 2011. doi: 10.1175/2011WCAS1061.1.

S. Fleming. *National Airspace System: DoT and FAA Actions Will Likely Have a Limited Effect on Reducing Delays During Summer 2008 Travel Season: Congressional Testimony.* DIANE Publishing Company, 2009. ISBN 9781437908244.

Yarin Gal. *Uncertainty in deep learning.* PhD thesis, PhD thesis, University of Cambridge, 2016.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in Neural Information Processing Systems*, 2017.

Ira Gershkof. Shaping the future of airline disruption management (IROPS). Commissioned by Amadeus, 2016.

Karthik Gopalakrishnan and Hamsa Balakrishnan. A comparative analysis of models for predicting delays in air traffic networks. In *Twelfth USA/Europe Air Traffic Management Research and Development Seminar (ATM2017)*, 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017.

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)*, pages 1–6, Sept 2016. doi: 10.1109/DASC.2016.7778092.