Flight Delay Analysis and Prediction

Yellow Jacket No Delay: Sihan Liu , Yi Xie, Tian Jin, Bangtian Zhou, Jiaxing Liang, Yaqin Song

Motivation

A **flight delay** is when an airline flight lands and/or takes off later than the time scheduled. In the United States, the Federal Aviation Administration estimates that flight delays cost airlines $22 billion each year. Flight delay is also an inconvenience to passengers by making them late to their personal scheduled commitments and events. A connecting flight could be missed for a passenger who is delayed on a multi-plane trip. Delayed passengers can be angry and frustrated, and are forced to rearrange the pre-scheduled event, which may incur extra cost on hotel booking, car rental, etc.

To solve these problems, this project is to build a model based on historical flight delay data using machine learning techniques to **predict future flight delays**. This is valuable for airlines, passengers and other associated business.

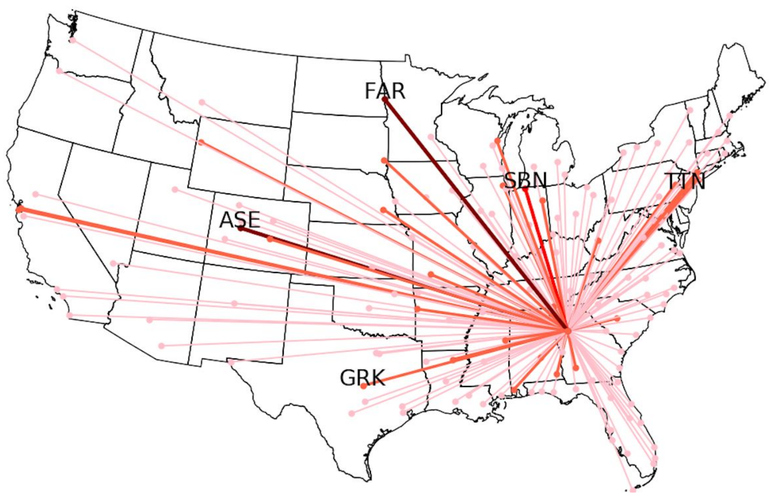


Figure 1: Flight delay visualization of flights departure from Atlanta, GA. Higher average departure delay is shown with deeper color.

Approach

We selected two machine learning algorithms for model building, and aggregate the results from two models to get final prediction.

Random Forest Regressor

Random Forest is an ensemble learning method for classification or regression that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

- Input factor value is transformed into multi-column '0, 1' data.
- Max_features is set to be √m
- n_estimators is set to be 100.
- Min_sample_leaf is tested and tuned to be 5000.

LSTM

Long Short Term Memory (LSTM) network is a special kind of Recurrent Neural Networks (RNN), capable of learning long-term dependencies. Since previous delays may have large impact on following flights' departure time, we decide to use RNN for delay prediction, which takes into account the impact of previous events. LSTM overcomes the vanishing gradient problem with traditional RNNs and can keep short-term memory for long period of time.

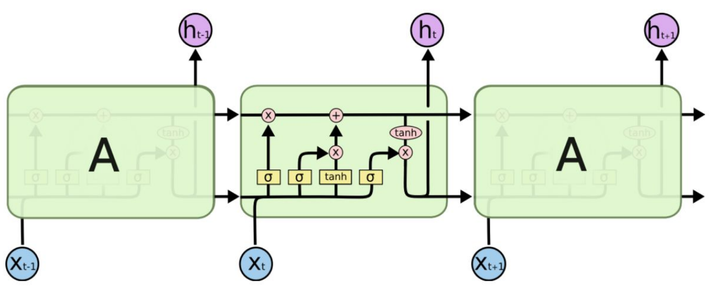


Figure 3[3]: An LSTM block is composed of four main components: a cell, an input gate, an output gate, and a forget gate. Gates can be viewed as regulators of the flow of values that goes through the connections of LSTM.

Model parameters:

- hidden layer neurons=30
- optimizer=rmsprop (RMSProp is recommended for RNN)
- epochs=20
- batch size=100

Flight Data Analysis

Before selecting features and making the prediction, we need to analysis the effect of each attributes on flight delay, or how flight delay is related with the values of each attributes. To do this, we utilize the data analysis tool **Tableau** for data visualization, hence study the correlation between each selected factor and flight delay time, and try to find out some patterns. Following are visualized result for average departure delay time respects to quarter, weekday, destination city, and carrier. (Figures are for 2016 data)

- Quarter 3 (month 7,8,9) seems to have higher average delay. Weather might be one of the reason, and people may tend to travel in summer more than in other seasons.
- Days before and after weekends (Thursday ,Friday, Monday) have higher delay, probably because more people fly for weekend vacation, causing more late-aircraft delays.
- For Destination City (Departure city is ATL), Fargo, ND has the highest delay, followed by Traverse City, MI.
- For carriers, Frontier Airlines (F9), Spirit Airlines (NK), and JetBlue (B6) have higher average delay than other carriers, probably because these are low-cost carrier.

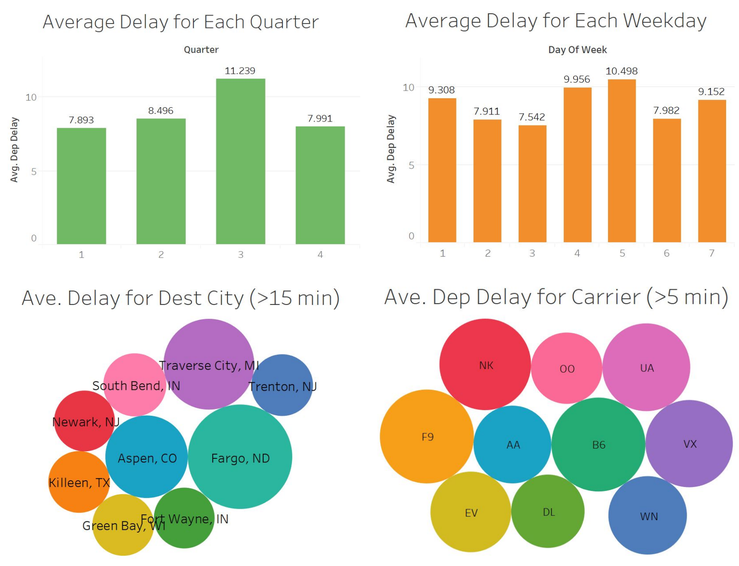
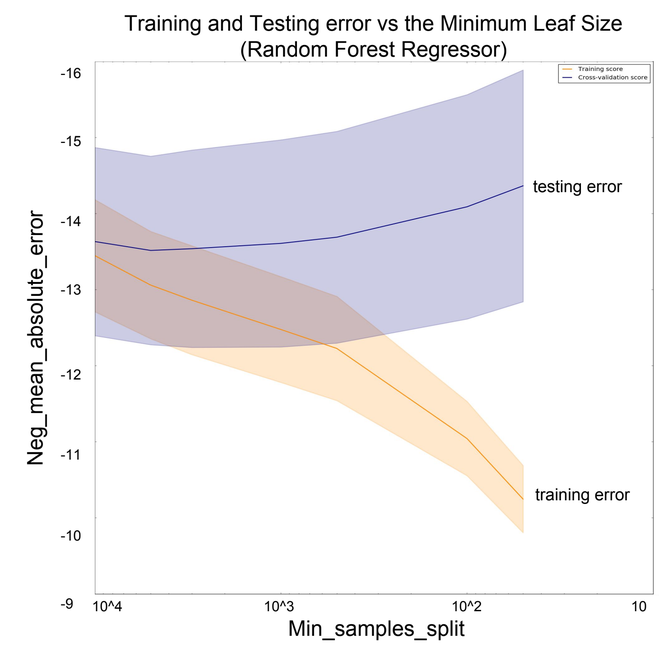


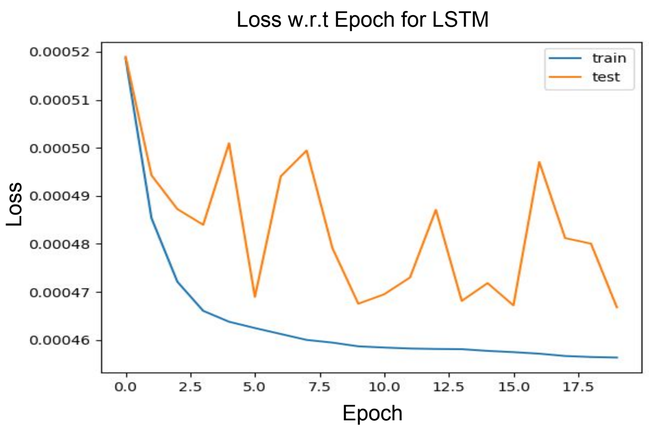
Figure 2: Average departure delay with respect to different attributes.

Result

Random Forest Regressor: As tuning the parameter of the Random Forest model, we can see as the min_sample_split goes smaller, the training error becomes smaller. However, after a certain point, the testing error stops decreasing. In the end, we picked the value that gives us the smallest testing error.

LSTM: The performance of LSTM on test data improves with epochs. However, we can see that there are fluctuations for testing data, probably because of overfitting of training data. We set learning rate to 0.001. Smaller learning rate leads to slower convergence and larger learning rate tends to produce high variance.





Data Preprocessing

We use flight data together with former flight info, holiday info and weather info to train the model.

- Flight Data were obtained from **United States Department of Transportation**[1]. Each flight data record all the parameters for each flight.
- Flight delay propagation were also considered base on the airplane tail number.
- Weather data are from **Aviation Weather Center**[2] and combined with flight data based on flight departure airport, destination airport and time. The most significant weather data such as precipitation, pressure altimeter, visibility, sky level 1 coverage and sky level 1 altitude were considered.

Sample data:

MONTH	DAY_OF_MONTH	DAY_OF_WEEK	UNIQUE_CARRIER	TAIL_NUM	FL_NUM	ORIGIN	DEST
1	1	5	AA	N157UW	2020	PHX	CLT

CRS_DEP_TIME	previous_delay	holiday	p01i	alti	vsby	skyc1	skyl1	p01i2	alti2	vsby2	skyc12	skyl12	DEP_DELAY
15	No previous	1	0	30	10	FEW	22000	0	30.13	10	SCT	12000	-3

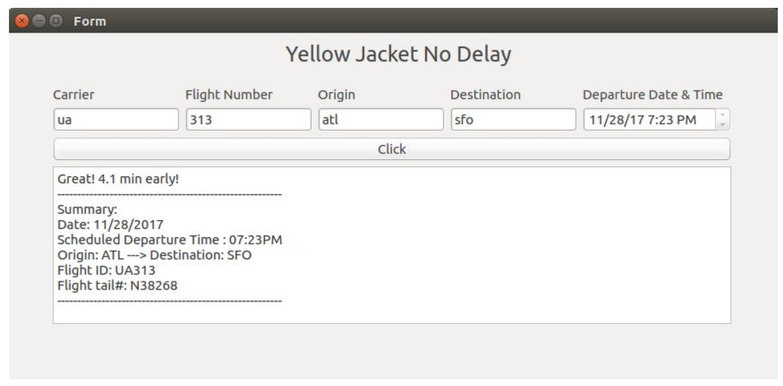
Out of all data entries, we select only flights **departure from ATL** (around 1 million records, 2015.01 - 2017.09), and our aim is to predict delay time of flights departure from ATL.

Real Time Prediction

We would like to test our model against real time flight data.

- Inputs including date and time, flight ident and tailnumber, origin and destination, departure time, previous delay in minute of the same aircraft, holiday, and weather parameters such as precipitation, pressure, visibility, and sky conditions for both airports are fed to our models and output the departure delay in minute.
- A gui acquires inputs from a user such as flight ident, origin, destination and departure time, then it acquires other inputs from flightware Api [4] and pymetar [5]. Results obtained by the trained LSTM model using these input will be shown on the gui interface.
- An example: from ATL to SFO, airline UA 313, schedule departure time 11/28/2017 7:23 pm.
  Result from our model: departed 4.1 minutes early
  Result from flightware: departed 7 minutes early

GUI Interface:



Reference

[1] Flight data source: **United States Department of Transportation**: https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

[2] Weather data source: **Aviation Weather Center** : https://mesonet.agron.iastate.edu/request/download.phtml

[3] C. Olah. "Understanding LSTM Networks", August 27, 2015 http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[4] FlightXML API: https://flightaware.com/commercial/flightxml/

[5] Pymetar: http://www.schwarzvogel.de/software/pymetar.html