A Novel Approach: Airline Delay Prediction Using Machine Learning

Swaminathan Meenakshisundaram
Assistant System Engineer
TCS Limited, Chennai, India
swaminathan.m1@tcs.com

Shubham Sinha
Data Scientist
TCS Innovation Labs - IITMRP
Chennai, India
shubham.sinha1@tcs.com

Gautham Balasubramanian
Data Scientist
TCS Innovation Labs - IITMRP
Chennai, India
gautham.balasubramanian@tcs.com

Natarajan Vijayarangan
Senior Scientist
TCS Innovation Labs - IITMRP
Chennai, India
n.vijayarangan@tcs.com

Abstract— **Flight delays often create exasperation in airports when not properly mobilized. Increasing development of machine learning models provoke the researchers and scientists to harness the modern research problems. Due to an increase in customer satisfaction in the air transportation system, there needs a proper decision-making process to mobilize the air-traffic with minimum delay. It is recorded that 19% of United States domestic flights reach their destination with an average delay of 15 minutes. Moreover, the complexity of the air transportation system limits the availability of accurate prediction models. Due to the stochastic nature of delays, this research investigates the qualitative prediction of airline delays to implement necessary changes and provide better customer experience. Collection of historical weather data and operational data during departure and arrival at airports serves as the source for building prediction models. A logistic regression model is used to get the status of the delay which is further contrasted to a decision tree model for evaluating the performance of the delay. The proposed research empirically evaluates the effectiveness of the decision tree algorithm over logistic regression. The results of this simulation indicate the potential delays in major airports including the time, day, weather, etc., and hence the volume of delay shall be minimum based on the constructed model.**

Keywords— Flight delay prediction, Logistic regression, Decision tree algorithm, Analytical modelling, Delay evaluation.

I. INTRODUCTION

Air transportation system is one of the crucial modes of modern versatility. With increasing congestion in air-traffic and passenger-traffic, it is important to maintain persistence and resilience [4]. Availability of land and resources contribute to the infrastructure of airports. The norms of improving technology and procedure are to maintain safety, efficiency, capacity, etc., Therefore, the National Airspace System (NAS) focuses on minimizing the environmental effects as a result of improvisation. With the current technology in hand, passengers can visualize their flight path, altitude, heading and other related parameters during their journey. However, air-traffic authorities continuously try to depreciate the delay in departure and arrival of flights. Though their efforts were in phase, the outcome is undesirable as the delays are in terms of hours sometimes causing chaos. Some important parameters that cause delay include weather, maintenance, security, and carrier. Corporate travel and tourism are the two major contributors to flight transportation system which is expected to be doubled by 2030. As a result of this increase, the air-traffic is also expected to increase in the same multiple. To minimize the air-traffic congestion new airports can be constructed. But, the complexity still grows exponentially. Hence, the only possible way of minimizing the delay is to improvise the existing airports. Considering the limited availability of land resources, the latter is more of a logical solution. Delay basically represents the period by which the aircraft is late or cancelled. Commercial aviation is likely to be affected if there is a delay in their mobility. This delay results in the dissatisfaction of trusted customers and sometimes even marketing strategies. With a view of understanding the flight system, scientists and researchers stored the vast amount of data recorded over the entire course of a flight journey.

In Brazil, almost 16.3 % of flights (including domestic and international) suffered a delay of nearly 30 minutes during the year 2013. With the vast amount of data recorded from sensors and IoT, giant organizations involved their data scientists for searching every potential information that can reduce the delay. Delays are usually classified into two categories namely the departure delay and arrival delay. Timely forecasting of flight delay could help aviation administrators to take necessary actions thereby reducing the economic loss. The standard aviation procedures allow the passenger to choose the right itinerary and flight of their own choice. Also, the aviation board is responsible for regulating the duties of pilots, attendants, cleaners, etc., This routine happens through the year. It is estimated that an achieving flight delay prediction model could save money in terms of millions. As an example, United States saves $ 1.6 million based on an appropriate approach [5].

As per the amendments from ANAC [1], it is formulated that the airline is responsible for minimizing the discomfort of passengers. Some of them include informing the passengers about the delay periodically, providing them with refreshments when needed, etc., So, the researchers started integrating modern techniques with the airport sectors to make them much more intelligent than before.

Hence, this paper progresses with the design of data science models to abridge the path between the passengers and flights.

978-1-7281-1360-9/18/$31.00 ©2018 Crown
DOI 10.1109/CSCI46756.2018.00210

This project is carried out based on the publicly available data from the Bureau of Transportation Statistics.

II. LITERATURE REVIEW

Since flight delays cause numerous problems across the globe, there has been a significant improvement in delay prediction models right from the late 90's. The quantity of the delay lessens the quality of marketing strategies. A delay in the departure or arrival of a domestic flight affects the operation of an international flight. A small amount of change in the delay value can be a massive amount of success for airport sectors. The first and the foremost thing is the availability of airline data. Due to the data deluge, the Federal Aviation Administration (FAA) and other aviation boards lease the information publicly to support a massive amount of research across the globe.

'Delay Multiplier' is a concept developed by Beatty in 1998 [2] which was proved to be the stepping stone for flight delay predictions. This theory represents the after effect of a delayed flight. The initial delay is used to estimate the collective delay of all the flights connected together. This research was carried out with Berry and Rome at Oak Ridge National Laboratory. Their findings indicate that a small change in the root value has a huge impact. However, their theory is well short for changing scheduling strategies. After this point, there are numerous outcomes witnessed. One such theory was developed by Cohn [7] at the Massachusetts Institute of Technology along with his colleagues. They investigated the delay propagation in the United States focusing only on two major airlines. Their findings indicate that the resulting delay is in the order of several hours where the impact is not very severe. It also investigates the number of passengers involved, the number of crew members changed, etc., However, their findings do not account for the other flights.

Later due to the popularity of machine learning, the research shifted towards the building of statistical predictive models. Statistics was found to be very useful when applied to real-world problems as it is purely analytical. Artificial Neural Networks (ANN) is another powerful technique that finds applications ranging from medicine to engineering. The recent developments in neural networks constitute the Recurrent Neural Networks (RNN). A deep learning based flight delay prediction approach [10] is carried out using the RNN. The popularity of Bayesian networks forced the research to apply the theory to flight delay predictions. A custom flight delay estimation based on Bayesian networks was studied by Xu et al [9]. There are a variety of studies focusing on weather-related parameters. Another method of estimating the flight delay based on probability distributions is given by Tu et al [8].

Centre Office for Delay Analysis (CODA) is responsible for the collection of operational airline data for further delay processing. This organization investigates the operational parameters in real-time during the pursuit of the journey. In cooperation with the Bureau of Transportation Statistics (BTS), this paper builds a logistic regression model and a decision tree model for the observed flight operational data. Finally, the results are empirically evaluated against each other

III. PROBLEM STATEMENT

The main objective here is to utilize the available flight operational data and data mining techniques to construct an analytical model. The analytical model constructed here is used to predict the flight delay based on some of the flight attributes which will be discussed in the latter section of this paper. Additional models will be created to determine the most likely cause of a flight delay and to predict the approximate duration of the delay.

IV. RESEARCH ATTRIBUTES

The flight delay is estimated based on the following attributes ORIGIN, DESTINATION, DATE_OF_JOURNEY, TAXI_IN, TAXI_OUT, WEATHER_CONDITIONS, etc., The data for this project is taken from BTS where an on-time performance database contains flight information including scheduled and actual arrival times as reported by the United States commercial airlines in Comma Separated Values (CSV) formatted data files. Therefore, the entire list of attributes used in this project are listed below:

1. YEAR
2. ORIGIN
3. DESTINATION
4. CRS_DEP_TIME
5. CRS_ARR_TIME
6. ARR_TIME
7. ACTUAL_ELAPSED_TIME
8. DISTANCE
9. AIRPORT
10. TAXI_OUT
11. TAXI_IN
12. DAY_OF_WEEK
13. TAIL_NUM
14. FL_NUM
15. DEP_TIME
17. CRS_ELAPSED_TIME
18. AIR_TIME
19. QUARTER
20. UNIQUE_CARRIER

V. DELAY PREDICTION METHODOLOGY

Based on the attributes listed above, this project yields the following outcome attributes which signifies the slot for the mobility of aircrafts

1. ARRSLOT
2. DEPSLOT

It is found that the flights are characterized into on-time and delayed from the observed dataset in BTS. The flight delays are predicted based on the arrival attribute. Many kinds of research are going on in the prediction of flight delay based on departure attribute as well. The arrival attribute contains the total number of delay minutes relative to the scheduled arrival

time. The dataset also contains five delay category attributes that quantify the reason for the delay. For the purpose of brevity, they are explained as follows

CARRIER_DELAY (CD): Delay due to the carrier. (Mechanical & Administrative)
LATE_AIRCRAFT_DELAY (LAD): Delay in the arrival of the scheduled airline.
NAS_DELAY (NASD): Delay due to air traffic.
WEATHER_DELAY (WD): Delay due to weather conditions.
SECURITY_DELAY (SD): Delay due to security problems.

Each of these fields contains the total number of delay minutes for each category. Since a flight can be delayed for multiple reasons, arrival delay is the summation of the five delay categories as shown below

$$\text{Delay}_{AD} = \sum \text{Delay}_{(CD + LAD + NASD + WD + SD)} \quad (1)$$

Apart from the primary flight details, general weather parameters including precipitation, humidity, and visibility contributing to natural phenomena such as thunderstorms, hailstorms, and overcast weather at the origin and destination airports have been considered additionally. A raw dataset often consists of missing values (usually represented as NaN). It is not advisable to build a model with missing values as it will not project the desired outcome. So, to treat the missing values in the dataset obtained from BTS, data pre-processing is an initial step.

The mean values of the corresponding airports are used for filling the missing values in TAXI_IN and TAXI_OUT attribute. Since the dataset consists of multiple columns, a record consisting of more than four missing values is said to be filtered. The data also contains units with numbers which has to be removed before building the model. Hence the dataset is successfully filtered by restricting the UNIQUE_CARRIER to be UNITED_AIRLINES (UA) thereby also limiting the number of airports to ten.

TABLE I
AIRPORTS CONSIDERED FOR DELAY PREDICTION

Airport Name	Type of Operation	IATA
Newark Liberty	International	EWR
Logan	International	BOS
Los Angeles	International	LAX
San Francisco	International	SFO
George Bush	Intercontinental	IAH
O'Hare	International	ORD
Denver	International	DEN
LaGuardia	International	LGA
Orlando	International	MCO
Ronald Regan Washington	National	DCA

After all these initial steps, the analytical model is built using logistic regression and decision tree. A comparison is also made between the obtained results.

Logistic regression is one of the most populous classification algorithms that work based on the results of a confusion matrix. A simple linear regression is given below

$$y = \theta_0 + \theta_1 x_1 \quad (2)$$

θ_0 and θ_1 represent the bias and weight of the parameter x_1. The simple linear regression model has a linear relationship between the dependent and independent variables x and y. Hence, the logistic regression is applied as a binomial function to the simple linear regression. The formula for the logistic regression is shown below by modifying the equation (2)

$$P(x_1, x_2) = 1 / (1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}) \quad (3)$$

The represented equation computes the probability based on a Sigmoid function: $1 / (1 + e^{-z})$. For "z" as an input into the function, we include a linear multiplication of the parameters θ and x, where

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \quad (4)$$

The above formula is used to get the status of the delay which is later improvised to retrieve the quantity of delay. The logistic regression is in the form of a Sigmoid function as shown below

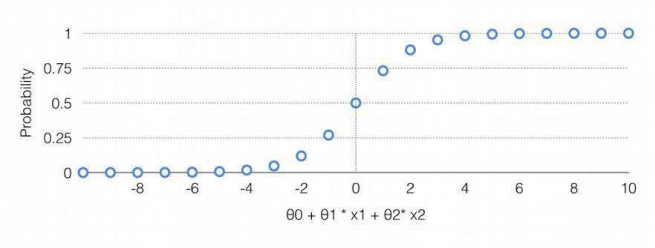

Fig. 1. Probability distribution of the logistic regression model.

Therefore, the logistic regression yields a Boolean result which lies in the likelihood distribution between 0 and 1. The advantage of this model is that it says whether the delay is there or not (usually discrete "Positive" or "Negative"); rather it does not represent the magnitude of delay. Indication of the presence of delay is not really ideal to minimize the problems mentioned in the introduction and survey.

Since the output is a probable value that lies between 0 and 1, there is a need for a conversion technique to convert these probability values into classifications. This is usually done by using the decision boundaries.

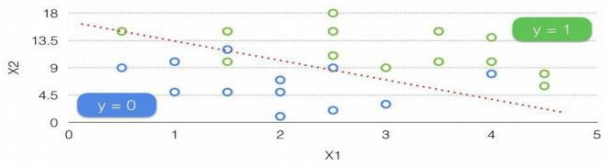

Fig. 2. Classification of binary outcomes with a decision boundary.

A decision boundary is the line which is used to separate the input examples, and thereby designating the classification of positive class as (y = 1) and negative class as (y = 0). Now after applying the boundary, the accuracy of boundaries should be measured.

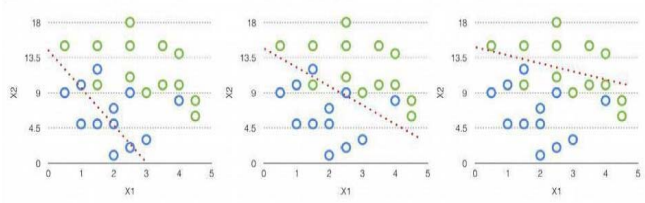

Fig. 3. Accuracy of decision boundaries

The best boundary can be fixed only by varying the parameters θ_0, θ_1 and θ_2. After the successful variation in parameters, the best boundary available with the best accuracy can be sorted out for appropriate classification algorithm. Hence, to measure the accuracy of logistic regression, the following equation with cost function is utilized.

Cost Function and Gradient Descent Algorithm

A cost function is a measure of performance of a prediction model. It estimates the error in the predicted values based on the relationship between the input and output parameters. Logistic regression equation for finding the Estimated Time of Arrival / Departure (ETA / ETD) is given below:

$$\text{ETA / ETD} = m_1 x_1 + m_2 x_2 + m_3 x_3 + \ldots + m_s x_s \quad (5)$$

The parameters (m_1, m_2, m_3, …, m_s) are the gradient values (i.e., slope or weight) and the parameters (x_1, x_2, x_3, …., x_s) are the flight attributes. For instance, x_1 = CRS_DEP_TIME, x_2 = TAXI_IN, x_3 = TAXI_OUT, etc.,) Since the cost function is calculated based on the Mean Squared Error (MSE), the equation describing the cost function used for calculating the error is shown below:

$$J(\text{ETA / ETD}, x_a) = \frac{1}{N} \sum_{i=1}^{n} (x_a - (m_1 x_1 + m_2 x_2 + m_3 x_3 + \ldots + m_s x_s))^2 \quad (6)$$

The above equation can be simplified as follows:

$$J(\text{ETA / ETD}, x_a) = \frac{1}{N} \sum_{i=1}^{n} (x_a - (\text{ETA / ETD}))^2 \quad (7)$$

where 'N' is the total number of data recorded in the observation. The estimated delay values are compared with the actual delay values recorded in the BTS dataset. The estimated cost function's magnitude is descended to the minimum optimal value using the Gradient Descent algorithm as follows:

$$\begin{aligned} f'(m_1) &= -x_1 (x_a - (m_1 x_1 + m_2 x_2 + m_3 x_3 + \ldots + m_m x_m)) \\ &\vdots \\ f'(m_s) &= -x_s (x_a - (m_1 x_1 + m_2 x_2 + m_3 x_3 + \ldots + m_s x_s)) \end{aligned} \quad (8)$$

$f'(m_1)$, $f'(m_2)$, $f'(m_3)$, …., $f'(m_s)$ are the weights obtained using Gradient Descent algorithm. These weights are multiplied and subtracted from the learning rate to find the steepest ASCENT. The following equation specifies the operation mentioned above:

$$m_s = m_s - (\text{Learning\_Rate} * \text{mean}(f'(m_s))) \quad (9)$$

This subtraction returns the weight values that are used to minimize the cost function. Therefore, it is advisable to train the model with a larger iteration number. It is found that at more than 1000 iterations, the error seems to be considerably reduced. The weights refer to the coefficient of parameters in these machine learning models.

Now first we will run our model with some initial weights & then gradient descent updates our weights. And try to find the closer cost value after going through the no of iteration.

After finding the results from logistic regression, this project now proposes the decision tree process for predicting the airline delay. Random forest machine learning model is flexible at most of the times as it does not require a hyper-parameter tuning. Thereby yielding satisfactory results most of the time. It is the most often used model due to its simplicity and in fact it can be used for both classification and regression applications [6]. Combination of multiple decision trees yield a better result in random forest model.

A flowchart describing the flow of operation on the flight data through the described models are shown below

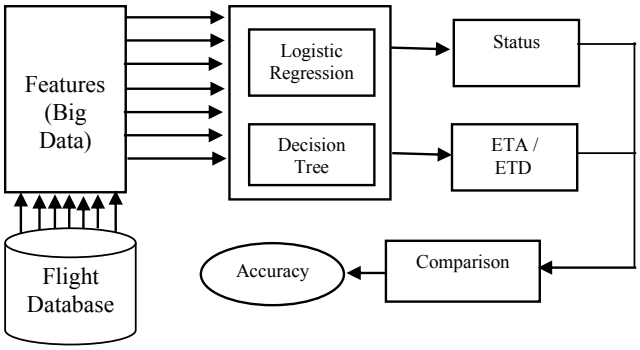

Fig. 4. Operational flow for the flight delay prediction

ETA represents the estimated time of arrival. Based on the above flow, the flight parameters from BTS database are taken as features and fed into the machine learning models. The models are compared with each other and the accuracy of each model is outlined in the latter section of this paper.

VI. OBSERVATIONS

For the purpose of brevity, this project has chosen only ten airports in the United States of America. The quantity of arrival and departure delay based on the above constructed models are observed below

TABLE II
DEPARTURE DELAY ACCURACY FOR AIRPORTS

IATA Code	Delay Accuracy
DCA	0.8892773892773893
BOS	0.7721164613661814
DEN	0.815587603559374
EWR	0.7932213044716605
IAH	0.7798387096774193
LAX	0.8909221385001794
LGA	0.8367496339677891
MCO	0.8034906270200388
ORD	0.7861515727253536
SFO	0.798372513562387

TABLE III
ARRIVAL DELAY ACCURACY FOR AIRPORTS

IATA Code	Delay Accuracy
DCA	0.8352941176470589
BOS	0.7551020408163265
DEN	0.8485516372795969
EWR	0.7543302540415704
IAH	0.8588298443370908
LAX	0.8388660947407683
LGA	0.7766116941529235
MCO	0.8415193189259987
ORD	0.8218537414965986
SFO	0.8033573141486811

It is observed that the resulting outcome in terms of accuracy from decision tree is better than the logistic regression model. Hence, the overall delay for both arrival and departure of flights obtained by both the models are compared below

TABLE IV
DELAY ACCURACY BY BOTH MODELS

Delay Type	Logistic Regression	Decision Tree (Random Forest)
Arrival	0.8168	0.8812
Departure	0.8083	0.8489

Hence, to focus the objective, a decision tree algorithm is modeled which gives the magnitude of delay in terms of accuracy.

According to [3], it is found that the results obtained from decision tree algorithm has better accuracy than the logistic regression model. Factor analysis is used to understand the possible factors affecting the delay of a flight. Hence, the analyzed factors are implemented using the random forest algorithm. The dataset is divided in the ratio of 4:1 for both training and testing sets. The results from both the machine learning models are subjected to the confusion matrix to understand the number of false positives and false negatives.

To find the accuracy of the given dataset through these models, the following procedure is taken into consideration

1. The dataset is run using various models and obtain the delay predictions.
2. For each prediction, the output is compared for different models.
3. The accuracy of each outcome is found with true values.
4. The intersection of correct predictions is found for each model.
5. The intersection found is the true accuracy.

The models are executed for different iterations in the set (k_1, k_2, …, k_m). Out of these executions, 'n' are predicted correctly. The 'n' predictions can be obtained only by using the Gradient Descent algorithm. The predicted values are evaluated against the actual values thereby which the accuracy of each machine learning model can be found. The accuracy is improved by increasing the number of iterations in Gradient Descent slope.

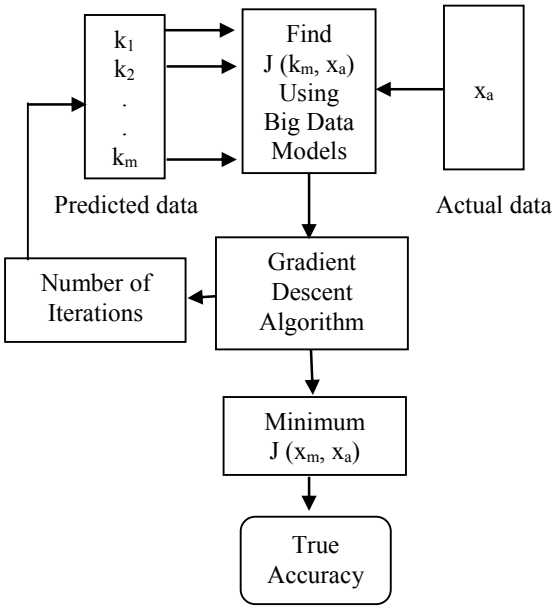

Fig. 5. Operational flow for finding true accuracy

x_a represents the actual value and x_m represents the measure value. From the above diagram, it is clear that the final true accuracy depends on the number of iterations referring to Fig. 4. During the first run, we get x_1 set of predicted values of delay. This set is now compared with the actual data stored in the testing set. The cost function J is used to find the error variations between the actual and predicted delay. Lesser the error, better the accuracy.

So, the magnitude of cost function is reduced down the slope using Gradient Descent Algorithm by increasing the number of iterations. Finally, the true accuracy represents the best machine learning model in terms of its predicted values that match likely with the actual data. Therefore, the inference obtained as a result of the above simulation is shown below

TABLE V
INFERENCES

Sl. No	Inference
1	More delay in Fridays and Mondays
2	More departure delays between 4 – 8 PM
3	More arrival delays between 7 – 10 PM
4	Boston airport has more arrival and departure delays than any other airports
5	More delays in the second and third quarter of the year where it depends on seasonality
6	Flights to and from EWR and BOS are found to have more delays
7	DCA – EWR airports are found to have more delays
8	High temperature caused more delays
9	Low visibility caused more delays
10	Los Angeles airport is found to be the one with less quantity of delay when contrasted to others

VII. CONCLUSION AND EXTENSION

In this detailed study, flight attributes recorded in BTS are used to predict the status of the delay and also the length of the delay. Several non-flight attributes including weather and peak travel data are also invoked for this study thereby to yield a meaningful result. Since the original dataset is categorized into training and testing sets, the resulting outcome is contrasted to the testing set thereby evaluating the efficiency of models. Hence, the outcome measured produces almost 80 – 85 % true accuracy for both the machine learning models.

The commercial aviation system is complex in which an event occurring in one of the occasions may cause ripple effects in other areas. Consequently, a model solely relying on the flight attribute will fail when accounted for complex interactions across the system. As an extension, this research may to choose to include the additional assets of data by employing deep learning methodologies. Firstly, system congestion and airport congestion include airport capacity metrics and scheduled flight data. The information contained here may be relevant to estimate the system and air-traffic congestion to predict the quantum of delay. Secondly, the age of the aircraft may be taken into consideration for building the machine learning model. A brief information about this age can lead to predict the likelihood of a hardware maintenance delay.

VIII. REFERENCES

[1] ANAC. Agencia Nacional de Aviac¸ ao Civil. Technical report, http://www.anac.gov.br/ 2017

[2] Beatty, R., Hsu, R., Berry, L. and Rome, J.: Preliminary evaluation of flight delay propagation through an airline schedule. In 2nd USA/Europe air traffic management r&d seminar, Orlando, 1.-4.12.1998.

[3] Gopalakrishnan, Karthik, and Hamsa Balakrishnan, "A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks", 2016.

[4] K. R. Chandramouleeswaran and H. T. Tran, "Data- driven resilience quantification of the US Air transportation network," Annual IEEE International Systems Conference (SysCon*)*, Vancouver, BC, 2018, pp. 1-7, 2018.

[5] Lawson, D. and Castillo, W. Predicting flight delays. Technical report, Computer Science Department, CS 229, Stanford University, Stanford, CA, 2012.

[6] Rebollo de la Bandera, and Juan José, "Characterization and Prediction of Air Traffic Delays", DSpace@MIT, Massachusetts Institute of Technology, 2012, http://dspace.mit.edu/handle/1721.1/76107

[7] Shervin AhmadBeygi, Amy Cohn, and Yihan Guan, "Analysis of the Potential for Delay Propagation in Passenger Aviation Flight Networks", April 1, 2007.

[8] Tu, Y., Ball, M. O., and Jank, W. S. "Estimating flight departure delay distributions - A statistical approach with long-term trend and short-term pattern", Journal of the American Statistical Association, 103(481), pp. 112–125, 2008.

[9] Xu, N., Laskey, K.B., Donohue, G., Chen, C.H. "Estimation of delay propagation in the national aviation system using Bayesian networks", In: 6th USA/Europe Air Traffic Management Research and Development Seminar, 2005.

[10] Y. J. Kim, S. Choi, S. Briceno and D. Mavris, "A deep learning approach to flight delay prediction," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, pp. 1-6, 2016.