The Prediction of Flight Delay: Big Data-driven Machine Learning Approach

Jiage Huo[1], K. L. Keung[1], C. K. M. Lee[1], Kam K. H. Ng[2], K.C. Li[1]
[1]Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[2]Interdisciplinary Division of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, China
(jiage.huo@connect.polyu.hk, dicky-kin-lok.keung@connect.polyu.hk, ckm.lee@polyu.edu.hk, kamhung.ng@ntu.edu.sg, ryan.kc.li@connect.polyu.hk)

Abstract - **Nowadays, Hong Kong International Airport faces the issues of saturation and overload. The difficulties of selecting taxiways and reducing the lead time at the runway holding position are the severe consequences that appeared from increasing the number of passengers and increased cargo movement to Hong Kong International Airport but without constructing a new runway. This paper is primarily about predicting flight delays by using machine learning methodologies. The prediction results of several machine learning approaches are compared and analyzed thoroughly by using real data from the Hong Kong International Airport. The findings and recommendations from this paper are valuable to the aviation and insurance industries. Better planning of the airport system can be established through predicting flight delays.**

Keywords – **Big Data, Flight Delay, Prediction, Machine Learning**

I. INTRODUCTION

With the increasing demand for air transport, the pressure on the planning and controlling of airlines has increased significantly. Nowadays, in the airline system, flight delay is the most crucial prevalent operation indicator showing the performance of an airline. Any discrepancy between the scheduled and the actual departure or landing times of an aircraft can be seen as a delay [1].

Besides focusing on the Hong Kong aviation's data, flight delays are prevalent in most countries. Governmental Regulatory Authorities have various criteria concerning flight delays tolerance thresholds. In reality, in the environment of air transport systems, flight delay is a vital topic. In 2013, 36% of flights in Europe were delayed over five minutes, 31.1% of flights in the United States were delayed over 15 minutes, and 16.3% of flights in Brazil were canceled or delayed over 30 minutes. The results indicate how important this metric is and how it impacts, regardless of the airline mesh size [2].

Flight delays may weaken the performance of the transportation system, and will bring adverse impact on the planning of an airport. Airlines are liable to sanctions, fines and additional costs caused by flight delays [3]. Thus, airlines need to predict flight delays, and understand their causes as well as having better reactions. The flight delay calculation will foster airports and airline managers' tactical and operational decisions and inform passengers of rescheduling their plans. Researchers establish forecasting models to rectify the root delay, predicting the time, location, reasons and sources of the delay. This comprises models effectively looking for an estimation of the probability, the number of minutes, or the level of delay for a particular flight or airline [4, 5]. Ionescu, et al. [6] showed the recurrent failure of flight schemes to respond to unexpected delays, technological failures, airport congestion and air traffic late coverage also resulted in problems of current data analysis.

Big Data has earned wide attention over the past few years. The accessibility of Big Data has offered a new generation of technology to different industries, shifting commercial attention towards data-driven decision making. Big Data Analytics is presently an essential component of gathering Business Intelligence. Many businesses, especially massive enterprises, consider Big Data as mainstream practice. They are continually investigating the latest tools and models in order to improve their Big Data utilization.

Flight delays can be triggered by various sources and greatly impact airports, airlines or on-the-way routes. Some airport schemes incorporated airlines and destinations as the factors into the data analysis to predict flight delays. Airports also have top priority of investigation on the efficiency based on the delays caused by carriers. The linkage between congestion and delay has been examined in an ensemble of the airport and the route. Other airports and airlines need to be considered for evaluating the capacity problems and airlines' decision making. There are many ways to analyze the entire scope, and it is important to consider the dynamics of air transport systems, especially if root delays are targeted [7].

This study aims at predicting flight delay with different approaches. Machine learning approaches provide opportunities for the prediction of flight delays, and gain increasingly more attention. In order to find a good method for flight delay prediction, several machine learning methods are introduced and applied in predicting flight delays. The prediction results with different machine learning methods are compared so that the best prediction model can be chosen. The flights dataset used in this research was obtained from the Hong Kong International Airport Authority. Meanwhile, feature variables are developed and selected before they are used in the prediction process. **Table I** shows the corresponding references and research objectives for the 10 methods used in predicting flight delays.

978-1-5386-7220-4/20/$31.00 ©2020 IEEE

TABLE I
A SUMMARY TABLE FOR METHODS USED IN PREDICTING FLIGHT DELAYS

Methodology	Reference to Related studies	Objective
Network representation	[8]	Investigation of the flight structure
Probabilistic model	[9]	Distribution of delays (If delays are less than 2 hours, the probability can be predicted.)
Statistical analysis	[10]	Comprehend delay propagation impacts
Queuing model	[11]	Predict root delays
Simulation	[12]	Analyze airport capacity data, considering departure and arrival delays
Operational model	[13]	Assess the flight delayed cost of the airline schedule
	[14]	Study the impact of queuing capabilities and demand on airport delay levels
	[15]	Study the spread of delays
Random forests	[16]	Flight delays prediction
Reinforcement learning	[17]	Temporary taxi-out
Naive Bayes, decision tree, neural network and K-nearest Neighbours	[18]	Flight delays prediction (Decision tree performs best with a 70 percent prediction on longevity.)
Adaptive network	[19]	Flight delays prediction

II. METHODOLOGY

Hong Kong's airline data in 2018 was obtained from the Hong Kong Airport Authority databases. The dataset includes the scheduled and actual departure and arrival times for non-stop flights recorded by different airlines. Information for the delayed and cancelled flights, real travelling time and non-stop distance are also available in the dataset. The origin and destination of airlines are also included. With this dataset, a predictive model was used to process flight delays. The flight dataset includes data for 161 airports. A flight arriving later than the scheduled arrival time is considered as a flight delay. In addition, the diverted flight is eliminated from the dataset. Features such as a month, day of the week, origin airport code, destination airport code, departure time, arrival time are used. Unnecessary features are removed from the flight data.

This study aims to predict flight delays for airlines in Hong Kong. Five methods are used to predict flight delay, that is, Random Forest, Logistics Regression, K-nearest Neighbors, Decision Tree, Naïve Bayes.

A. Random Forest

Random forest is one of the algorithms used in this study. It is a term that falls within the general random decision technique. This algorithm works by creating a group of decision-making structures in training and outputting the class, which is the class mode or the mean prediction of each of the tree. The random selection of attributes is used for the determination of splits at each node [20].

The value of the decision tree for each function is determined in the following way:

$$f_{i_i} = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} n_{i_j}}{\sum_{k \in all\ nodes} n_{i_k}} \quad (1)$$

At the Random Forest point, the last feature is that it is an average for all the trees. The value of the function is determined and divided by the total number of trees:

$$RFf_{i_i} = \frac{\sum_{j \in alltrees} normf_{i_{ij}}}{T} \quad (2)$$

B. Logistics Regression

Regression is a statistical model that uses a logistic function in its basic form to predict a binary dependent variable even if there are several more complex extensions. Logistic regression is used to estimate logistic model parameters in a regression analysis. In mathematics, a binary logistic model has two possible values, for example, pass or fail, represented by a variable indicator that the values 0 and 1 are defined as two values. The standard logistic function is expressed below:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}} \quad (3)$$

The term t can be represented as follows:

$$t = \beta_0 + \beta_1 \chi \quad (4)$$

and the general logistic regression function can subsequently be expressed as:

$$p(\chi) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \chi)}} \quad (5)$$

C. K-nearest Neighbors

K-nearest neighbors is a simple algorithm that holds all the available cases in a similarity measure and classifies new cases. A case is defined by a plurality of its neighbors, with the case assigned by the distance to the most common of its k neighbors. If $k = 1$, the case is only assigned to the nearest neighbor's class. The distance functions of the k-nearest neighbors are indicated below:

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \quad (6)$$

$$\sum_{i=1}^{k} |x_i - y_i| \quad (7)$$

$$\left(\sum_{i=1}^{k} (|x_i - y_i|)^q \right)^{\frac{1}{q}} \quad (8)$$

D. Decision Tree

A decision tree is a structural model for supervised learning, applying the strategy of dividing and occupying. At every decision node providing terminal leaves, the data is split into smaller subsets, using different conditions. At each n decision node, a $f(x)$ test function is implemented to offer the branches with discrete results. According to the

input variable x, the mathematical classification model of the decision tree can be defined as:

$$f(x) = E[y|x] = \sum_{n=1}^{N} w_n \phi(x; v_n) \quad (9)$$

where w_n is spread over class labels in the nth node. v_n is the option of the split variable and the threshold value on the root to the n^{th} leaf path.

E. Naïve Bayes

The class mark is distributed according to the probability obtained from the Bayes theorem. For convenience, the variables are assumed to be independent, and the variables and groups follow a normal Gaussian probability curve [21].

III. DISCUSSION

Deep understanding of feature selection is highly beneficial in enhancing model efficiency and knowledge of the structure and features of the data. The following variables were used to forecast delays by different classification models:

(1) Month. It is the month of one year (1-12), and provides information on seasonality.
(2) Day of week. It ranges from 1 (Monday) to 7 (Sunday), and offers regular flight patterns and results.
(3) Departing and arrival times (per hour). It offers patterns based on the flights' time. For example, flights arriving late at night could be less likely to delay.
(4) Distance. It is the distance in miles between airports of origin and destination.
(5) Airport origin or destination. It means the airport code of origin, such as HKG for the Hong Kong International Airport.
(6) Carrier. It is the business code of an airline and the operating flight number. For instance, CX is the code that is used for the Cathay Pacific.

Then, additional features such as average taxi time, airline group, block time, etc. are incorporated into the further analysis to understand how these attributes impact on the accuracy of algorithms and the flight delay issue.

The dataset was split into training and test data in a 80:20 ratio, within 29,352 flight data points, and a sampling technique was used to balance the training results. The efficacy of the models was assessed according to precision. The configuration of the computation unit used was Anaconda Jupyter Notebook, Python 3.7 version. Apart from data processing and feature selection, different libraries were imported in Jupyter Notebook, including Pandas, Seaborn, Sklearn and Sklearn.matrix.

Table II shows the comparison results of the algorithms' accuracy after testing different additional attributes, namely average taxi time (the average of taxi in and taxi out time), airline group, block time, distance (distance in miles between airports of origin and destination) and aircraft group size. After adding these extra attributes for analysis, they primarily increase the accuracy of several of the algorithms. The detailed results are as follows. The accuracy of random forest lies between approximately 67% to 73%, compared to the original accuracy of 66.39%. K-nearest neighbors has accuracy that lies between approximately 63% to 66% compared to the original accuracy of 61.03%. Finally, the accuracy of the Naïve Bayes lies between 49% to 65% compared to the original accuracy of 49.17%. Naïve Bayes has the maximum increase in its accuracy. In contrast, the additional attributes decrease the accuracy of the logistic regression and decision tree. However, the decreases in these two algorithms are not as large as the increases in the other algorithms. The new accuracy of the logistic regression and decision tree is still quite near the original accuracy. As a result, the above additional attributes have the most impact on Naïve Bayes, then random forest, and ultimately k-nearest neighbors, but do not significantly influence logistic regression and decision tree. Regarding the effects of each attribute, after comparing the altered extent in each accuracy by each attribute, distance is the most crucial attribute, block time is the second one, aircraft group size is the third one, and airline group is the fourth one. The final one is the average taxi time.

TABLE II
THE ACCURACY OF ALGORITHMS BY EACH ATTRIBUTE

Algorithm	**Random forest (Original accuracy: 66.39%)**				
Extra attribute(s)	Average taxi time	Airline group	Block time	Dist.	Aircraft group size
Accuracy	69.58%	70.41%	72.80%	67.19%	69.95%
Algorithm	**Logistic regression (Original accuracy: 62.39%)**				
Extra attribute(s)	Average taxi time	Airline group	Block time	Dist.	Aircraft group size
Accuracy	62.51%	62.17%	62.31%	62.37%	62.17%
Algorithm	**K-nearest neighbors (Original accuracy: 61.03%)**				
Extra attribute(s)	Average taxi time	Airline group	Block time	Dist.	Aircraft group size
Accuracy	63.15%	64.25%	63.33%	65.58%	63.69%
Algorithm	**Decision tree (Original accuracy: 61.74%)**				
Extra attribute(s)	Average taxi time	Airline group	Block time	Dist.	Aircraft group size
Accuracy	60.71%	60.71%	60.71%	61.11%	61.34%
Algorithm	**Naïve Bayes (Original accuracy: 49.17%)**				
Extra attribute(s)	Average taxi time	Airline group	Block time	Dist.	Aircraft group size
Accuracy	49.17%	49.55%	52.17%	64.80%	52.04%

Since airlines and airports can make use of Big Data for analytics, they can refer to the above machine learning approach for improving the flight delay issue. Airports can then further examine this area for improvement. Block time, indeed, means the time between when the flight starts leaving the gateway, the actual flight duration, to the time when the flight arrives at the destination gateway. So, apart from the flying duration, the time and planning of the

airport' runway and gateway are also the vital components of the block time. By understanding the above implications and insights brought about by the block time attribute, airports can subsequently undertake further adjustments in this area to cope with the flight delay issue.

After conducting the above project analysis, the approach used in predicting flight delays can also be applied in the insurance industry. This project suggests that airlines can provide flights related data to insurance companies through using the application programming interface (API) technology so that insurance companies can obtain different flight related attributes for travel insurance products analysis. In fact, there are various factors involved in determining the pricing and coverage of travelling insurance products by different insurance companies. In the past, admittedly, insurance companies would also include different factors to analyze insurance products. However, they were not able to obtain flights related data to conduct analysis. By implementing the data sharing between airlines and insurance companies, insurance companies can obtain more types of data to allow a more comprehensive analysis regarding the flight delays\ prediction. Thus, insurance companies can then obtain a better understanding of the flight delay probability and determine better pricing and coverage of their travel insurance products.

The above data sharing process between airlines and insurance companies can be achieved by using an application programming interface (API). It is a software intermediary that permits two applications to communicate with each other. In other words, an API is the messenger who sends your request to the provider who returns the answer to you (Pearlman, 2016). By exchanging airline and flight related data through an API, airlines can collaborate with insurance companies to develop applications and services in predicting flight delays.

There are two proposed data sharing system workflows. The first one is to establish a connection between airlines and insurance companies only. The second one is to cooperate with the Hong Kong Airport Authority. A platform working as the middle stage for storing flight information provided by airlines is established. On the other hand, insurance companies can also extract the data from this official platform for further analysis.

IV. CONCLUSIONS

This research aims to investigate the classification of flights as on-time and delayed by using various machine learning algorithms and variables related to airlines and airports. Different variables influencing the performance of a flight are included. According to the results, the random forest approach has the highest accuracy when predicting flight delays.

In the future, the prediction of the flight delay will be used in the management of airport and airlines. Besides, other approaches will also be tested in future research to boost accuracy. Furthermore, additional factors related to airlines and external variables like weather will be explored in order to increase the prediction precision of the models.

ACKNOWLEDGMENT

This work was supported in part by the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong, in part by the Interdisciplinary Division of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. Our gratitude is also extended to the Research Committee and the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China, and The Innovation and Technology Commission, The Government of the Hong Kong SAR, Hong Kong for support of this project (PRP/002/19FX/K.ZM31). Our gratitude is also extended to the Research Committee and the Interdisciplinary Division of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University for support of the project (BE3V). The authors would like to express their appreciation to the Hong Kong International Airport and FlightGlobal for their assistance with the data collection.

REFERENCES

[1] K. K. H. Ng, C. K. M. Lee, and F. T. S. Chan, "An Alternative Path Modelling Method for Air Traffic Flow Problem in Near Terminal Control Area," in *2019 2nd International Conference on Intelligent Autonomous Systems (ICoIAS)*, 28 Feb.-2 March 2019, pp. 171-174, doi: 10.1109/ICoIAS.2019.00037.

[2] K. K. H. Ng, C. K. M. Lee, S. Z. Zhang, K. Wu, and W. Ho, "A multiple colonies artificial bee colony algorithm for a capacitated vehicle routing problem and re-routing strategies under time-dependent traffic congestion," *Computers & Industrial Engineering,* vol. 109, pp. 151-168, 2017, doi: https://doi.org/10.1016/j.cie.2017.05.004.

[3] R. Britto, M. Dresner, and A. Voltes, "The impact of flight delays on passenger demand and societal welfare," *Transportation Research Part E: Logistics and Transportation Review,* vol. 48, no. 2, pp. 460-469, 2012, doi: https://doi.org/10.1016/j.tre.2011.10.009.

[4] K. K. H. Ng, C. K. M. Lee, F. T. S. Chan, and Y. Qin, "Robust aircraft sequencing and scheduling problem with arrival/departure delay using the min-max regret approach," *Transportation Research Part E: Logistics and Transportation Review,* vol. 106, pp. 115-136, 2017, doi: https://doi.org/10.1016/j.tre.2017.08.006.

[5] K. K. H. Ng, C. K. M. Lee, F. T. S. Chan, C.-H. Chen, and Y. Qin, "A two-stage robust optimisation for terminal traffic flow problem," *Applied Soft Computing,* vol. 89, p. 106048, 2020, doi: https://doi.org/10.1016/j.asoc.2019.106048.

[6] L. Ionescu, C. Gwiggner, and N. Kliewer, "Data Analysis of Delays in Airline Networks," *Business & Information Systems Engineering,* vol. 58, no. 2, pp. 119-133, 2016, doi: 10.1007/s12599-015-0391-3.

[7] K. K. H. Ng, C. K. M. Lee, F. T. S. Chan, and Y. Lv, "Review on meta-heuristics approaches for airside operation research," *Applied Soft Computing,* vol. 66, pp. 104-133, 2018, doi: https://doi.org/10.1016/j.asoc.2018.02.013.

[8] K. F. Abdelghany, S. S. Shah, S. Raina, and A. F. Abdelghany, "A model for projecting flight delays during irregular operation conditions," *Journal of Air Transport Management,* vol. 10, no. 6, pp. 385-394, 2004, doi: https://doi.org/10.1016/j.jairtraman.2004.06.008.

[9] Y. Tu, M. O. Ball, and W. S. Jank, "Estimating Flight Departure Delay Distributions—A Statistical Approach With Long-Term Trend and Short-Term Pattern," *Journal of the American Statistical Association,* vol. 103, no. 481, pp. 112-125, 2008, doi: 10.1198/016214507000000257.

[10] D. Markovic, T. Hauf, P. Röhner, and U. Spehr, "A statistical study of the weather impact on punctuality at Frankfurt Airport," *Meteorological Applications,* vol. 15, no. 2, pp. 293-303, 2008, doi: 10.1002/met.74.

[11] F. Wieland, "Limits to growth: results from the detailed policy assessment tool [air traffic congestion]," in *16th DASC. AIAA/IEEE Digital Avionics Systems Conference. Reflections to the Future. Proceedings*, vol. 2: IEEE, pp. 9.2-1.

[12] G. Hunter, B. Boisvert, and K. Ramamoorthy, "Advanced national airspace traffic flow management simulation experiments and vlidation," in *2007 Winter Simulation Conference*, 9-12 Dec. 2007, pp. 1261-1267, doi: 10.1109/WSC.2007.4419730.

[13] M. J. Soomer and G. J. Franx, "Scheduling aircraft landings using airlines' preferences," *European Journal of Operational Research,* vol. 190, no. 1, pp. 277-291, 2008, doi: https://doi.org/10.1016/j.ejor.2007.06.017.

[14] A. M. Kim, "The impacts of changing flight demands and throughput performance on airport delays through the Great Recession," *Transportation Research Part A: Policy and Practice,* vol. 86, pp. 19-34, 2016, doi: https://doi.org/10.1016/j.tra.2016.02.001.

[15] N. Pyrgiotis, K. M. Malone, and A. Odoni, "Modelling delay propagation within an airport network," *Transportation Research Part C: Emerging Technologies,* vol. 27, pp. 60-75, 2013, doi: https://doi.org/10.1016/j.trc.2011.05.017.

[16] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," *Transportation Research Part C: Emerging Technologies,* vol. 44, pp. 231-241, 2014, doi: https://doi.org/10.1016/j.trc.2014.04.007.

[17] P. Balakrishna, R. Ganesan, and L. Sherry, "Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures," *Transportation Research Part C: Emerging Technologies,* vol. 18, no. 6, pp. 950-962, 2010, doi: https://doi.org/10.1016/j.trc.2010.03.003.

[18] L. Zonglei, W. Jiandong, and Z. Guansheng, "A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning," in *2008 International Symposium on Knowledge Acquisition and Modeling*, 21-22 Dec. 2008, pp. 589-592, doi: 10.1109/KAM.2008.18.

[19] S. Khanmohammadi, C. Chou, H. W. Lewis, and D. Elias, "A systems approach for scheduling aircraft landings in JFK airport," in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 6-11 July 2014, pp. 1578-1585, doi: 10.1109/FUZZ-IEEE.2014.6891588.

[20] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[21] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Machine Learning,* vol. 29, no. 2, pp. 103-130, 1997, doi: 10.1023/A:1007413511361.