2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT) | 978-1-6654-2518-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCASIT53235.2021.9633476

Research on Flight Delay Prediction Based on Random Forest

Peng Hu
The Second Research Institute of Civil Aviation Administration of China
Chengdu, China
hupengbaby@163.com

Jianping Zhang*
The Second Research Institute of Civil Aviation Administration of China
Chengdu, China
zhangjp@caacsri.com

Ning Li
Civil Aviation Administration of China
Beijing, China
lining@caac.gov.cn

***Abstract*—Based on the random forest model, this paper proposes a flight delay prediction model. By analyzing the departure flight data of Guangzhou Baiyun International Airport in June 2020, and selecting the data of ten landing airports, it analyzes the distribution of delayed, punctual, and early arrived. It studies the selection of features that impact on flight delays, and establishes random forest predictions model. Through case study, it researches the mean square error of different leaf sizes when the forest scale is 50 trees. The results show that the optimal leaf size is 5, and the minimum mean square error is 0.1096. And it analyzes the importance of features such as departure flight delay time, scheduled flight time, number of scheduled departure flights on the day, date, and landing airport. The research results also found that, when the forest size is 100 trees and the leaf size is 5, the out-of-bag mean square error is 0.1090, and the accuracy of the prediction model is high, which is close to 90%.**

Keywords—random forest, regression model, prediction model, flight delay, data analysis

I. INTRODUCTION

Flight delay has always been the focus and hot spot of social concern. With the rapid development of civil aviation, the number of flights continues to increase. Flight delays due to uncertain factors such as weather and flow restrictions have increased, and safety risks have increased, impacting on passenger travel. Therefore, predicting the regularity of flights can provide decision-making support for airport management of aircraft operations and provide decision-making basis for passenger travel. It has great significance to the high-quality development of civil aviation.

In recent years, the research trend of flight delay prediction is mainly divided into deep learning, machine learning, and simulation. Reference [1] proposed a new learning framework to predict flight departure delay time and arrival delay time in advance. Reference [2] proposed a new deep belief network to predict flight delays. Reference [3] considered the spread of flight delays and proposed a new layered integrated machine learning model to predict consecutive flights delay. Reference [4] proposed a method for predicting flight arrival delays based on Bayesian networks. Reference [5] proposed a new automated data-driven method to predict flight arrival time. Reference [6] considered the characteristics of flight delay periodic fluctuation, established simulation models based on RBF neural network, BP neural network and wavelet neural network to predict flight delays and compared the prediction accuracy under different neural networks to verify the effectiveness of the prediction. Reference [7] proposed an XGBoost prediction algorithm based on nonlinear weighting with considering the imbalanced characteristics of flight delay data, and verified the effectiveness of the algorithm's prediction through analysis of examples. Reference [8] built a Hadoop Distribute File big data platform based on randomly connected CliqueNet and considered weather effects, established a flight delay prediction model, and verified the effectiveness of the prediction model through case analysis.

This paper proposes to establish a regression prediction model based on the random forest. Through extracting and analyzing the flight data of Guangzhou Baiyun International Airport in June 2020, it researches on how to selecting features. With analyzing the case study, it studies the mean square error under different leaf sizes, and determined the optimal leaf size and forest scale of the random forest model. By analyzing the importance of each feature and the out-of-bag mean square error, the effectiveness of the flight prediction model is studied.

II. DATA DESCRIPTION

The flight data in this article is from June 1st to June 30th, 2020. There are a total of 12,998 flights departing from Guangzhou Baiyun International Airport(GBIA). Among them, the data from Guangzhou to ten other cities in China are selected for research. There are 4947 flights, of which Part of the original flight data is shown in Table I. The flight data features include: flight number, date, planned departure time, actual departure time, planned arrival time, actual arrival time, scheduled departures at GBIA on the day, and departure airport code, arrival airport code. According to Tab. I, the statistics on flights departing from Guangzhou to ten other cities are shown in Tab. II.

The flight delay in Tab. II refers to the flight whose actual arrival time is delayed by more than 5 minutes, the early arrived flight refers to the flight that arrives more than 5 minutes earlier, and the punctual flight refers to the flight that is delayed or less than 5 minutes in advance (including 5 minutes). It can be seen from Tab. II that in June 2020, the number of flight delays from GBIA to the ten airports was 3271, which was accounting for 66.12% of it. The number of flights arriving on time was 408, which was accounting for 8.25% of it. And the number of flights arriving early was 1,268, which was accounting for 25.63% of it. There are 210 flights from GBIA to Wuhan Tianhe International Airport, which of the 180 flights were delayed, and the number of delayed flights accounted for as high as 85.71% of it. For the flights from GBIA to Xiamen Gaoqi International Airport,

the number of punctual flights accounted for the highest proportion, which was 11.62%. The 661 flights from GBIA to Chongqing Jiangbei International Airport, which of the 235 flights arrived early, which accounted for 35.55% of it. The flights, departing from GBIA to these ten different airports in June 2020, have a relatively high proportion of flights delayed for more than 5 minutes. Establishing a model to predict flight departure delays will help scientifically plan flights and help civil aviation high-quality development.

TABLE I. RAW FLIGHT DATA

Flight No.	Date	Planned Departure Time	Actual Departure Time	Planned Arrival Time	Actual Arrival Time	Planned Departures on the day	Point of Departure	Point of Arrival
CDC8990	2020-06-13	16:15	16:32	18:30	18:07	446	ZGGG	ZSHC
CSN3463	2020-06-12	08:05	09:31	10:10	11:28	462	ZGGG	ZUCK
CSH9304	2020-06-12	09:50	10:07	12:10	11:59	462	ZGGG	ZSSS
CSZ9867	2020-06-13	17:05	17:39	19:20	19:29	446	ZGGG	ZSNJ
CCA1829	2020-06-13	16:15	16:23	18:40	18:14	446	ZGGG	ZSSS

TABLE II. FLIGHT STATISTICS

Point of Departure	Point of Arrival	Number of Delayed Flights	Number of Punctual Flights	Number of Early Arrival Flights	Number of Flights
Guangzhou Baiyun Airport (ZGGG)	Beijing Capital Airport (ZBAA)	320	42	114	476
	Changsha Huanghua Airport (ZGHA)	72	4	12	88
	Wuhan Tianhe Airport (ZHHH)	180	7	23	210
	Xiamen Gaoqi Airport (ZSAM)	143	23	32	198
	Hangzhou Xiaoshan Airport (ZSHC)	538	58	181	777
	Nanjing Lukou Airport (ZSNJ)	416	42	167	625
	Shanghai Pudong Airport (ZSPD)	108	6	30	144
	Shanghai Hongqiao Airport (ZSSS)	630	70	274	974
	Chongqing Jiangbei Airport (ZUCK)	357	69	235	661
	Chengdu Shuangliu Airport (ZUUU)	507	87	200	794
Total		3271	408	1268	4947

III. RANDOM FOREST MODEL

A. Model Overview

The random forest model[9, 10] is generally used to solve classification problems, and its characteristics are the sampling of replacement samples and the "double random strategy". Assume that the number of samples in the data set G is N, and each sample has K features. Extraction with replacement means that when a fixed number of samples are drawn from G, each sample needs to be returned to G. "Double random strategy" refers to randomly selecting samples and randomly selecting s ($s \le K$) features. The random forest model is as follows:

1) Construct training set T. It converts the feature data containing characters in the data set G into integer values instead. Samples are randomly selected from the data set, and returned to G after each extraction, and the number of samples drawn accumulates to N. A total of m rounds of sampling are performed, and the training set $T=\{t_1, t_2, \ldots, t_m\}$ is obtained.

2) Establish a decision tree D. It randomly selects s features for each sample in the training set T, which satisfies $s \le K$. A total of m rounds of random extraction of features are performed to form m independent decision trees $D=\{d_1, d_2, \ldots, d_m\}$.

3) Training. A vote is made for each decision tree, and the one with the most votes is the final output result, or the average of multiple decision trees is output as the final result.

B. Feature Selection

Feature 1. The departure flight delay time is determined by the difference between the scheduled departure time of the flight and the actual departure time, and it is a direct manifestation of the degree of flight departure delay. The value of the departure flight delay time is negative, which means the flight is delayed, and its value is positive, which means the flight takes off early. Feature 1 is an independent variable, and the calculation formula is as follows:

$$A_x^{ij} = PD_x^{ij} - RD_x^{ij} \quad (1)$$

In this formula, the A_x^{ij} represents the delay time of the departure flight x from airport i to j. The PD_x^{ij} represents the planned departure time of the flight x from airport i to j. The RD_x^{ij} represents the actual departure time of the flight x from airport i to j.

Obviously, when the departure flight delay time is a negative number, the larger the value, the greater the delay time, the higher the degree of impact on flight arrival, and the higher the possibility of arrival flight delay.

Feature 2. The planned flight time of a flight is the difference between the planned departure time of the flight and the planned arrival time. Feature 2 is an independent variable, and the calculation formula is as follows:

$$B_x^{ij} = PD_x^{ij} - PA_x^{ij} \quad (2)$$

In this formula, the B_x^{ij} represents the scheduled flight time of departing flight x from airport i to j. The PA_x^{ij} represents the scheduled arrival time of flight x from airport i to j.

The planned flight time of a flight is an estimate of the flight duration of the flight. It is calculated by the airport and the airline in advance based on the selected route path length and the average flight speed, and its value is relatively fixed.

Feature 3. The number of scheduled departure flights on the day *C* refers to the total number of scheduled departure flights that day. The higher the value, the denser the number of departure and departure flights. Due to the airport capacity and the capacity limitation of the control sector, and considering the possible impact of weather factors, if the flight that needs to take off and departs ahead is delayed, it will seriously impact on the subsequent flights that need to take off on the same day. Feature 3 is an independent variable.

Feature 4. Flight date *D* refers to the specific flight date of the flight. The number of scheduled departures for each day is different, and the number of scheduled flights to various cities is also different. Feature 4 is an independent variable.

Feature 5. Airplane landing airport *E* refers to the final destination airport of the flight. Feature 5 is an independent variable.

IV. Case Study

This paper selects departure flight delay time *A*, planned flight time *B*, the number of scheduled departure flights on the day *C*, date *D*, and aircraft arrival airport *E* as the independent variables of the random forest model. And the determination of whether the flight delayed *F* is used as the dependent variable. The specific formula is as follows:

$$F_x^{ij}=\begin{cases}-1 & RD_x^{ij}-RA_x^{ij}<-5\,\text{min}\\ 0 & -5\,\text{min}\leqslant RD_x^{ij}-RA_x^{ij}\leqslant 5\,\text{min}\\ 1 & RD_x^{ij}-RA_x^{ij}>5\,\text{min}\end{cases} \quad (3)$$

In this formula, the RA_x^{ij} represents the actual arrival time of flight *x* from airport *i* to *j*. The F_x^{ij}=-1 means the flight *x* from airport *i* to *j* is delayed. Correspondingly, the F_x^{ij}=0 means the flight arrives on time, and the F_x^{ij}=1 means the flight arrives early.

According to the selected independent variables and dependent variables, it uses Matlab's TreeBagger function to build a random forest regression model.

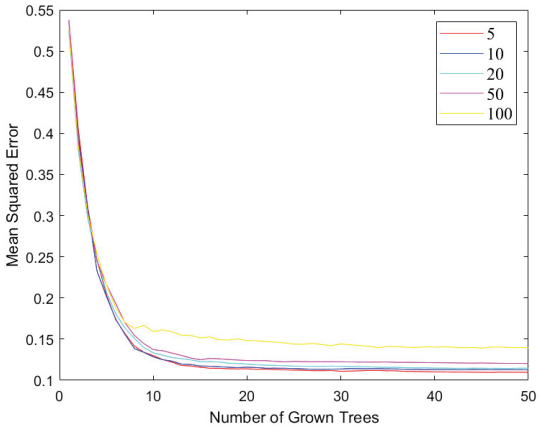

Fig. 1. The relationship between the number of grown trees and the mean square error with different leaf size.

As shown in Fig. 1, in order to reasonably determine the leaf size of each decision tree in the random forest, it compares the mean square error under different leaf sizes to determine the best leaf size as 5, and the mean square error is the lowest at this time, with a value of 0.1096.

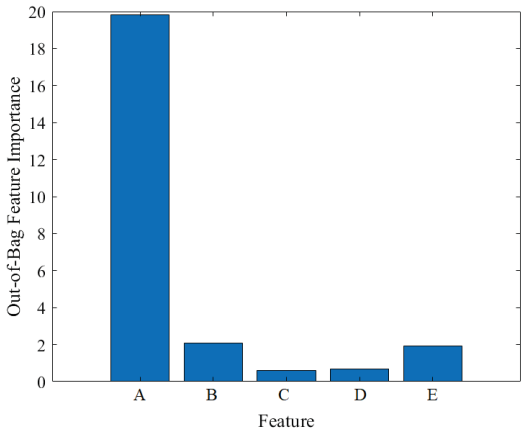

Fig. 2. Feature importance.

As shown in Fig. 2, according to the determined more reasonable leaf size 5, the importance of each feature is evaluated by using the growth scale of a forest of 100 trees. The larger the value, the greater the importance. It can be seen from the figure that the importance of features, and the departure flight delay time is the most important, followed by the planned flight time of the flight and the airport where the plane will land. The number of scheduled departure flights and the date of the day are less important.

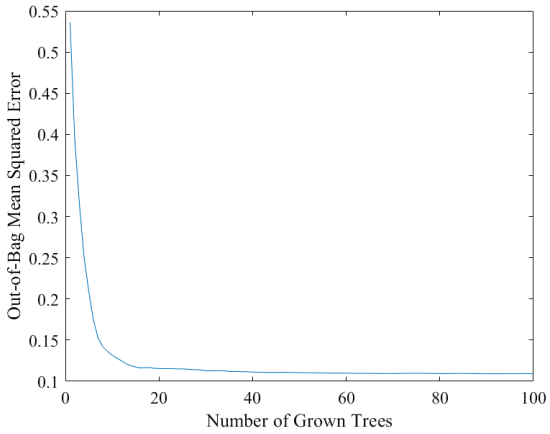

Fig. 3. The relationship between the number of grown trees and the out-of-bag mean square error.

As shown in Fig. 3, by using the leaf size of 5 and the forest size of 100, the minimum out-of-bag mean square error is 0.1090, and the prediction accuracy is 89.10%, which is close to 90%. And it means having a good prediction effect.

V. Conclusion

Through the case study section of calculation examples, the departure flight delay time, planned flight time and flight landing airport are important features, which are important factors impacting flight delays.

The random forest regression model adopts the optimal leaf size of 5 and the forest size of 100, the prediction accuracy is high, the mean square error is 0.1096, and the out-of-bag mean square error is 0.1090. The forecast

accuracy is close to 90%.

The random forest model established in this paper based on a variety of features has more accurate predictions for flight delays, punctuality, and early arrival. In the next research work, weather factors, airport capacity, airspace capacity, airport taxi time and other features can be considered. It accurately predicts flight delay, which could provide theoretical support for improving flight regularity, and help the high-quality development of civil aviation.

ACKNOWLEDGMENT

This work was partly supported by the Safety Foundation of Civil Aviation Administration of China ([2020] No.168) and partly by the Industry Colony Cooperative Innovation Project of Chengdu (2020-XT00-00001-GX).

REFERENCES

[1] J. Bao, Z. Yang and W. Zeng, "Graph to sequence learning with attention mechanism for network-wide multi-step-ahead flight delay prediction," Transportation Research Part C: Emerging Technologies, vol. 130, p. 103323, August 2021.

[2] B. Yu, Z. Guo, S. Asian, H. Wang, and G. Chen, "Flight delay prediction for commercial air transport: A deep learning approach," Transportation Research Part E: Logistics and Transportation Review, vol. 125, pp. 203-221, March 2019.

[3] W. A. Khan, H. Ma, S. Chung, and X. Wen, "Hierarchical integrated machine learning model for predicting flight departure delays and duration in series," Transportation Research Part C: Emerging Technologies, vol. 129, p. 103225, June 2021.

[4] Á. Rodríguez-Sanz, F. G. Comendador, R. A. Valdés, J. Pérez-Castán, R. B. Montes, and S. C. Serrano, "Assessment of airport arrival congestion and delay: Prediction and reliability," Transportation Research Part C: Emerging Technologies, vol. 98, pp. 255-283, December 2018.

[5] Z. Wang, M. Liang and D. Delahaye, "Automated data-driven prediction on aircraft Estimated Time of Arrival," Journal of Air Transport Management, vol. 88, p. 101840, July 2020.

[6] Q. F. Zhang, Y. Z. Wang, S. T. Wang, and K. X. Pei, "Prediction of flight departure delay time considering periodic fluctuation factors," Ship Electronic Engineering, vol. 41, pp. 133-136, July 2021.

[7] H. Tang, D. Wang, B. Song, W. K. Chu, and L. Y. He, "Classification of flight delay based on nonlinear weighted XGBoost," Journal of System Simulation,pp. 1-9, August 2020.

[8] J. Y. Qu, L. Cao, M. Chen, L. Dong, and Y. X. Cao, "CliqueNet flight delay prediction model based on clique random connection," Journal of Coumputer Applications, vol. 40, pp. 2420-2427, August 2020.

[9] L. Breiman, "Random Forests," Mach. Learn., vol. 45, pp. 5-32, October 2001.

[10] N. Meinshausen, "Quantile Regression Forests," J. Mach. Learn. Res., vol. 7, pp. 983-999, June 2006.