Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 1, 2025

Analysis of Aircraft Flight Safety Based on Random Forest Model and Its Broader Applications

Bradyn Turner

Department of Systems and Information Engineering, Virginia Tech, Blacksburg, USA

bradyn.turner49@vt.edu

Abstract: With the increasing importance of ensuring transportation safety, this study focuses on enhancing flight safety through data-driven methods. Given the limitations of traditional monitoring systems, which fail to identify the underlying causes of deviations in aircraft performance, this paper proposes an analysis framework leveraging flight data and random forest models. By pre-processing flight data and quantifying key influencing factors, the study identifies potential risks and establishes predictive models to minimize safety incidents. Furthermore, the versatility of the random forest model is highlighted, showing its applicability beyond aviation to domains such as healthcare, finance, and environmental monitoring, offering solutions for complex classification and regression problems.

Keywords: Flight Safety, Data-Driven Methods, Random Forest Models, Aircraft Performance Monitoring, Safety Incident Prevention

1. Introduction

With the rapid development of transportation industry, the safety of transportation operation has become one of the most concerned aspects of people's daily travel. As the safest and most convenient means of transportation, airplane has become a very important mode of transportation nowadays, so its flight safety cannot be ignored [1]. However, in reality, due to the weather, aircraft equipment and the pilot's own factors, the aircraft cannot always run smoothly, thus affecting flight safety [2]. At the same time, the traditional monitoring environment can only detect the abnormal flight data and correct the flight status of the aircraft, but there is no way to know the reason why the aircraft deviates from the normal operation status [3]. In this thesis, we analyze the existing flight data, set a series of parameters to study the inner mechanism and causes of aircraft overrun events, and summarize the normal data to establish a data analysis model; in this way, we can grasp the dynamic law of aircraft flight and predict the possible emergencies that may occur afterwards to minimize the occurrence of safety accidents [4].

2. Data Sources and Model Assumptions

The data in this article comes from question D of the mathorcup mathematical modeling challenge in 2023. In order to simplify the model and strive for its accuracy, the assumptions made in this paper are as follows: (i)The data found through the literature and the Internet are true and credible. (ii) There are no errors in variables such as time and date, and there are basically no errors in positioning variables with quantitative variables that have less data.

3. Construction and Solution of Random Forest-based Model

3.1 Data Pre-processing

(1)Missing value processing

By examining the data in the table, we find that the aircraft landing gear status data is partially blank, and after checking, we get that the blank aircraft landing gear data means the landing gear is retracted, then there is no abnormal value in it.

(2)Outlier processing

Through model assumptions, we exclude the possibility of outliers in the data of time, date, fixed class data and quantitative data with little data (see Appendix I). Because the data volume is too large, we first conduct reliability analysis to filter out the indicators with poor reliability of the data, and then process them for outliers.

3.2 Reliability Analysis

Coefficient of variation, also called "dispersion coefficient", can be used to describe the degree of dispersion of the results, and is a common statistic to measure the reliability of data. It is defined as the ratio of the standard deviation to the mean, and the formula is:

$$c_v = \frac{\sigma}{\mu}$$

Import the data, use SPSSPRO to perform descriptive statistics on the data, and determine the coefficient of variation of each indicator Among the various statistics, the variance and coefficient of variation are often used for reliability analysis, and a higher variance and coefficient of variation indicate that there may be outliers in the data, and here we set the coefficient of variation (CV) threshold to

0.15, and if the coefficient of variation of a variable is greater than this value, it indicates that the variable may not be reliable, with a higher probability of outliers, and we then process the outliers in it.

Let's take the variable PITCH ATT.2 as an example, the coefficient of variation of this variable is 0.454812, which is greater than 0.15, so there is a high probability of outliers in this variable.

3.3 Outlier Handling

Take TRA-L.1 and FMF GROSS WEIGHT as an example for outlier processing. we use the box plot method to determine the outliers, first draw the box plot of each indicator to find the outliers, Figure 1 and Figure 2 show the box plots of TRA-L.1 and FMF GROSS WEIGHT:

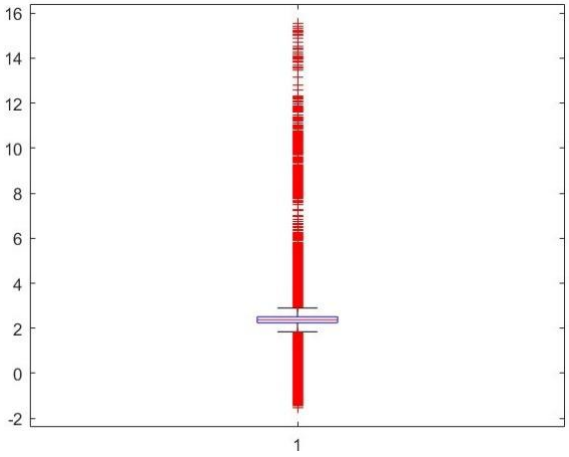

Figure 1. TRA-L.1 box diagram

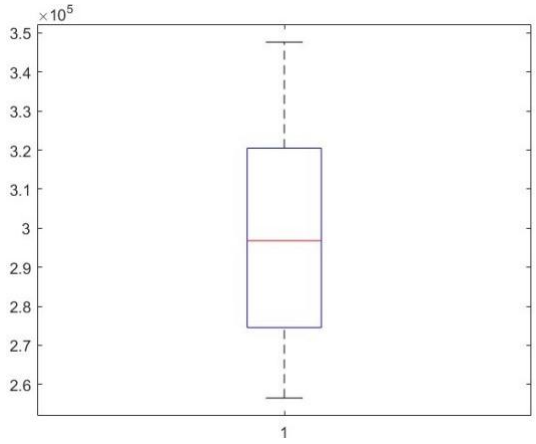

Figure 2. FMF GROSS WEIGHT box diagram

The outliers we find do not necessarily indicate data errors; it could also be a real data point. Therefore, the characteristics of the data itself and the actual situation need to be considered when identifying outliers. Through a rigorous consideration of data types and sources, we retained some of the outliers of the indicators.

Later, by observing the box plot outliers, the outliers were identified and removed, and the data preprocessing was completed by labeling the fixed class variables.

3.4 Random Forest based Weight Determination

If we want to find some data items related to flight safety, it is known from the title: G value is an important indicator to describe the safety at the moment of landing. We set the target quantity as G value, and extract the indicators with its strong correlation, and then give the weight of each indicator, which is known by the description of the key parameter segment data field: COG NORM ACCEL represents the G value of the aircraft landing, we take, the aircraft landing G value of 1 second, which is Q24-COG NORM ACCEL.9 as the dependent variable, and determine the importance of the remaining variables to the indicator degree [5].

Random Forest (RFF) is a machine learning algorithm that is an integrated model consisting of multiple decision trees. In Random Forest, each decision tree is constructed independently, and the output of each decision tree is voted to determine the final output [6].

The basic principle of the random forest algorithm is to perform classification or regression by constructing multiple decision trees on a training data set. Each decision tree is constructed based on a different random subset and a random set of features. This randomness allows each decision tree to have different features and samples, and reduces the risk of overfitting. When making predictions, the random forest votes on the predictions of each decision tree to determine the final output.

The analysis steps of the random forest algorithm are:

1.Import the data set into the random forest algorithm and determine the random forest model.

2.Use the random forest model to determine the weights, or levels of importance, of the variables.

3.Train the established model and test it using relevant data. The parameters of the obtained model are:

Table 1. Parameters of the model

Parameter Name	Parameter Value
Training time	18.142s
Data slicing	0.7
Data shuffling	No
Cross-validation	No
Node splitting evaluation criterion	MSE Maximum proportion of features considered for segmentation
Minimum number of samples for internal node splitting	2
Minimum number of samples in leaf nodes	1
Minimum weight of samples in leaf nodes	0
Maximum depth of the tree	10
Maximum number of leaf nodes	50
Threshold for impurity of node division	0
Number of decision trees	100
With put-back sampling true	TRUE
Out-of-bag data testing	FALSE

false	

In import we use the random forest model to determine the importance of each variable and determine them as weights to obtain the proportion of importance of each variable in the model as:

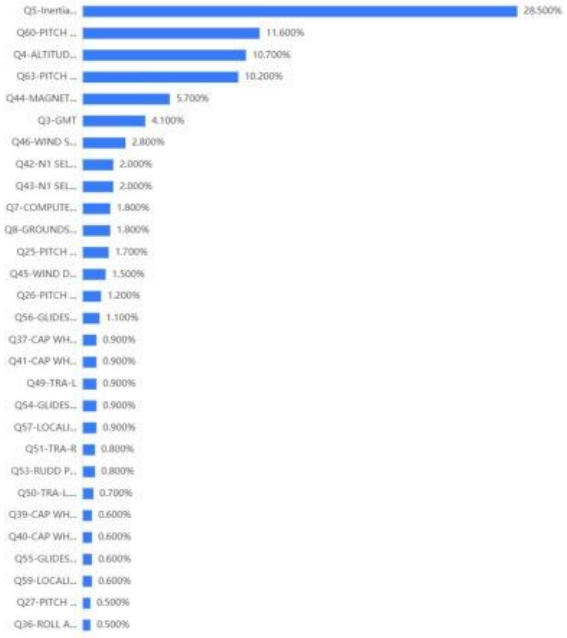

Figure 3. Importance ratio of each variable

The weights were determined by random forest model for longitudinal comparison.

Through the analysis of the importance ratio chart of each indicator, we can get: if we take the G value of landing 1s as the dependent variable and other indicators as the independent variables, the most important indicator for the flight safety of the aircraft is: Inertial Vertical Speed, i.e., descent rate, and through the search of relevant information, the descent rate is indeed an important indicator for the flight safety, which in side This confirms the correctness of the model.

The top five indicators in descending order of importance are as follows:

Inertial Vertical Speed (descent rate)

PITCH ATT RATE (pitch angle rate)

ALTITUDE (1013) (Altitude)

PITCH ATT RATE.1 (pitch angle rate)

MAGNETIC HEADING (magnetic heading)

As can be seen from the graph, the first five indicators account for 66.7% of the importance, so it can be considered that these five factors are the key data related to flight safety. We also determined the influence of these five indicators on aircraft flight safety by searching relevant information.

4. Quantitative Description Using BP Neural Network

4.1 Principle of BP Neural Network

BP neural network is a commonly used artificial learning network, often used to solve classification, regression, clustering and other problems. The algorithm has not only input layer nodes, output layer nodes, but also one to several implicit layer nodes. For the input signal, it is propagated forward to the implicit layer nodes, and then the output signal from the implicit layer nodes is propagated to the output nodes after the action function, and finally the result is output.

4.2 BP Neural Network Algorithm Steps

We build a 6-N-2 network mechanism, where the 6 represents the input quantity (respectively, the disc quantity, the rod quantity, the average value of the left and right hair throttle lever in one second and the average growth rate in one second), and the structure diagram is as follows:

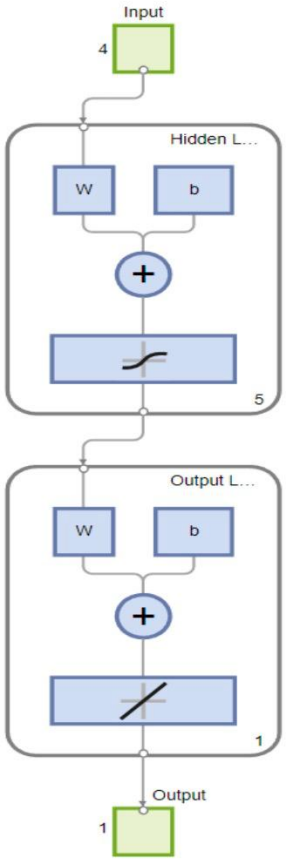

Figure 4. Structure diagram of BP neural network

(1)Initialize the weights: assign a set of small integer or negative values to wmi(0), wij(0), wjp(0) at random.

(2)Determine the structural parameters of the BP neural network and define the variables: Let Xk =[xk1, xk2,⋯, xkM], (k = 1,2, ⋯ , N) be the input vector, and N be the number of training samples. Yk(n) = [yk1(n), yk2(n), ⋯ , ykP(n)] is the actual output of the n iteration. dk = [dk1, dk2,⋯ , dkP] is the desired output.

(3)Input training samples: Input the training sample set X = [X1, X2,⋯, Xk,⋯ XN] in turn, so that the samples learned this time are Xk(k = 1,2, … , N).

(4)Forward propagation: Given the training pattern input, the actual output of the network is calculated and the training error of the samples is calculated.

(5)Back propagation: According to the error signal, calculate the error of the same unit, modify the weights and

thresholds. If the error is greater than K > N, go to step (6), otherwise go to step (3).

(6)Calculate the total error of network training, if it reaches the accuracy requirement, then end the training, otherwise go to step (3) and start a new round of training and learning.

4.3 Result Output

The variation of the model iteration error at the end of training is shown in Figure 5.

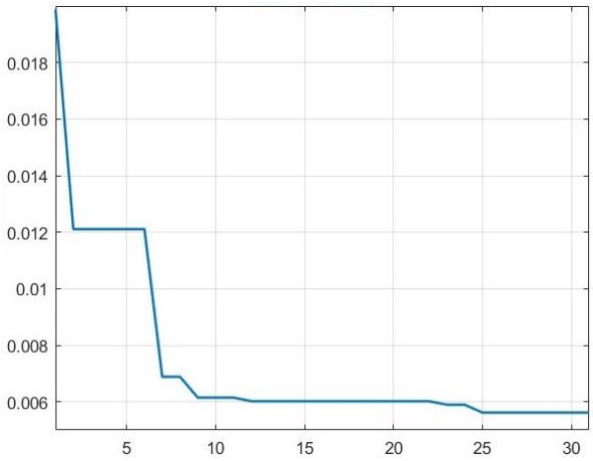

Figure 5. Variation of model iteration error

The trained neural network predicts the test data with RMSE=0.002363, which shows that the predicted output does not differ much from the expected output and basically achieves the desired effect

In summary, the quantitative description of the generation of the high point of G value and thus the low point of the pole position can be completed by using the pole volume, disc volume this and the position of the left (right) hair throttle lever to describe the quantitative description of the pole position as the cause of the occurrence of deviation.

The larger the value, the more likely the stick position will be significantly concave, the less smooth the pilot's maneuver, and conversely, the more normal the stick position will be and the smoother the pilot's maneuver will be.

5. Conclusion and Outlook

In this paper, we address issues related to aircraft flight safety by first pre-processing existing flight data and establishing specific quantification results of the importance of random forest for flight-related factors. In the future, the random forest model can be used to analyze many problems in medical, financial, and ecological fields, such as dealing with complicated genetic data, evaluating personal credit of bank accounts, studying the concentration of air pollutants and the pollution process, etc. The model can also be used for many problems in classification and regression. This model can also be used for many problems in classification regression.

References

[1] S. M. Hashemi, R. M. Botez, and G. Ghazi, "Robust trajectory prediction using random forest methodology application to UAS-S4 Ehécatl", Aerospace, vol. 11, no. 1, p. 49, 2024.

[2] W. Y. Yan and R. X. Gao, "Application of random forest to aircraft engine fault diagnosis", Proceedings of the 2007 IEEE International Conference on Automation Science and Engineering, pp. 1-6, 2007.

[3] J. K. Williams, "Using random forests to diagnose aviation turbulence", Machine Learning, vol. 92, no. 1, pp. 51-70, 2013.

[4] R. Patgiri, S. Hussain, and A. Nongmeikapam, "Empirical study on airline delay analysis and prediction", arXiv preprint arXiv:2002.10254, 2020.

[5] E. Fournier, S. Grihon, and T. Klein, "A case study: Influence of dimension reduction on regression trees-based algorithms - Predicting aeronautics loads of a derivative aircraft", arXiv preprint arXiv:1812.02310, 2018.