10.18686/aitr.v2i3.4413

Research on Factors Influencing Aviation Safety: Data Analysis and Key Findings

Zihan Zhao*

College of Mathematics and Statistics, Sichuan University of Science & Engineering, Zigong Sichuan 643000

Abstract: This research paper amalgamates a plethora of analytical tools to undertake an exhaustive investigation into the intricate facets influencing aviation safety. Through an in-depth exploration of pilots' situational awareness, feature importance ranking, and empirical analyses, it unveils the pivotal factors that shape flight safety outcomes. By harnessing the power of cutting-edge techniques such as XGBoost, Random Forest, and Forward Selection, the study meticulously processes and dissects flight safety data, thereby furnishing invaluable insights essential for fortifying aviation safety protocols. The implications of these findings are poised to offer substantial support and strategic guidance for safety management frameworks and pilot training initiatives within the aviation industry, thereby fostering a safer and more secure aviation environment for all stakeholders involved.

Keywords: Flight Safety; Random Forest; XGBoost; Forward Selection

Introduction

Aviation safety has always been of global concern, and its influences are complex and varied. From technological developments to human factors, all are shaping the face of aviation safety[1]-[3]. The aim of this investigation is to delve deeper into the impact of these factors on aviation safety and provide a deeper understanding to enhance the safety of air travel.

Traditional methods have always played an important role in this aircraft manoeuvring field, covering a wide range of techniques and engineering tools. Deng Yang used CART and QCA methods to analyse the factors influencing general aviation safety accidents and their grouping paths[4]; Yao Di et al. used the accident causation 2-4 model to analyse the factors influencing the safety of ventilation flights in conjunction with the operational characteristics of general aviation[5]; and Yue Qiming et al. used a comparative analysis method to study the pilot's situational awareness and its measurement method[6]. By combing and analysing the studies in the literature, it is possible to gain insights into how researchers have utilised traditional methods such as empirical models, and the challenges and limitations of these methods in solving problems.

In this paper, the dataset is first preprocessed to deal with the missing values and outliers in the original dataset using data analysis tools. XGBoost, Random Forest, and Forward Selection are used to rank the importance of the features to extract some key data items related to flight safety, and their importance is analysed to get an intuitive picture of the importance of each key data item.

1. Data pre-processing

1.1 Data reliability analysis

This paper uses SPSS to carry out descriptive statistics on the data in Annex I to analyse the coefficient of variation (CV) of each data, due to space constraints this paper only uses the data in the first table for example, and the rest of the results of the processing are shown in the "Appendix: Coefficient of Variation Table".

Table1 Table of coefficients of variation

Variable name	Sample size	maximum values	minimum value	Average value	standard deviation	Coefficient of variation
Altitude	38672	36027	119	30966	7896	0.2550
Descent rate	38672	2465	-362	-0.536	439.8	-820.5
Radio Altitude	38672	2043	-204	532.61	1138.	2.1371
Calculated airspeed	38672	324.5	30	283.87	50.40	0.1775
Ground speed	38672	503.5	0	422.26	89.83	0.2127

Table 1 demonstrates the results of descriptive statistics, including sample size, maximum values, etc., which are used to examine the quantitative data as a whole. Those that are outliers or more prominent are analysed separately, such as high mean values.

1.2 Outlier Handling

In Table 1, many data are observed to exhibit a left-skewed or right-skewed distribution. Particularly, the deviation of quantitative variables such as the left and right throttle lever positions (angles), lever volume, etc., is more pronounced. Therefore, this paper will replace outliers with the mean and exclude some variables. The latter part of the paper will no longer explore their relationship with flight safety. Additionally, due to the inability of the computer to recognize certain field names in the original dataset, they have been coded to ensure more accurate and precise processing. Here, the raw dataset is encoded as follows: attitude elevation is coded as 0, attitude pitch is coded as 1. Whether the A/P is activated on either side, the duration of air-to-ground gate openings, etc., are uniformly coded to indicate whether they are of the same data type. "Yes" is coded as 1 and "No" is coded as 0.

2. Model building

2.1 Feature Importance Ranking--XGBoost Boosting Algorithm

After processing the data with missing values, outliers, and feature construction steps, this paper uses spss to fit the data to the sample data through XGBoost algorithm by adopting the five-fold cross-validation method to get the relationship between each feature and user satisfaction. The main idea of XGBoost is to build a new model by Newton-Paphson method in the direction of gradient descent with the reduction of residuals of the previous model. XGBoost makes it more efficient to find the optimal solution of the model. The main idea of XGBoost is to build a new model by Newton-Paphson method in the direction of the gradient decrease of the residuals of the previous model. XGBoost makes it more efficient to find the optimal solution of model.

The XGBoost model can be represented in the following form, where we agree to denote the sum of the first t trees and denote the tth decision tree, and the model is defined as follows:

$$\mathrm{f_t(x)} = \sum_{\mathrm{t=1}}^{\mathrm{T}} \mathrm{h_t(x)} \tag{1}$$

Since the model is generated recursively, the model of the tth is formed by the tth - 1, which can be written as:

$$\mathrm{f_t(x)} = \mathrm{f_{t\text{-}1}(x)} + \mathrm{h_t(x)} \tag{2}$$

The tree to be added each time is the error of the previous tree summation:

$$\mathrm{r_{t,1}} = y_i - \mathrm{f_{m\text{-}1}}(x_i) \tag{3}$$

XGBoost was used to compute each feature, and the features were ranked one by one in descending order of importance based on the effectiveness of the model on the validation set in order to validate the previous analysis.

2.2 Feature Importance Scoring--Random Forest

There is low (or no) correlation between the individual Random Forest models, i.e. between the decision trees that make up the larger Random Forest model. Errors in individual decision trees do not affect the overall results in the right direction. Random forests allow for feature selection, replacement trials, and removal of redundant, poorly correlated features.

The CART algorithm is a commonly used method for binary classification, the decision tree generated by the CART algorithm is a binary tree, whereas the decision tree generated by the ID3 as well as C4.5 algorithms is a multinomial tree, and the binary tree model will be more efficient than the multinomial tree operation from the operational efficiency point of view. The CART algorithm selects the optimal features by means of the Gini (Gini) index.

In a classification problem, assuming that there are K classes and the probability that a sample point belongs to the kth class is, the Gini coefficient of the probability distribution is defined as:

$$Gini(p) = \sum_{k=1}^{K} p_k(1 - p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{4}$$

If CART is used for a two-class classification problem (not only for binary classification), then the Gini coefficient of the probability distribution can be simplified as:

$$Gini(p) = 2p(1 - p) \tag{5}$$

Let the Gini coefficient of the dataset D divided into two parts and D2 according to feature A when feature A is used to divide the dataset D into two parts and D2 be:

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{6}$$

2.3 Feature sequence selection--forward selection (SFS)

In order to solve the problems of Random Forest algorithm in high dimensional data sets, this paper chooses the forward selection (SFS) method. There may be some feature redundancy or feature irrelevance in high-dimensional data, and these factors will adversely affect the performance of the classifier, forming the so-called "dimensional catastrophe", which will inevitably increase the training time of the model and decrease the accuracy performance. Therefore, in order to avoid these problems, a feature sequence selection method is needed to remove redundancy and correlation between features and to reduce the dimensionality of the dataset. In the forward selection process, the optimal features that are highly correlated with the classification performance are filtered from the original features to improve the accuracy of the algorithm. This not only improves the performance of the original algorithm, but also reduces the time consumption needed to train the model. Therefore, forward selection is an effective feature selection method that can improve the performance of the algorithm while reducing the algorithm running time, and is suitable for high-latitude datasets, and the basic steps are shown in Fig.1:

In this paper, we adopt one of the sequential forward selection methods to select 9 feature variables from the original dataset of many feature variables, and plot the marginal effect of feature importance as follows:

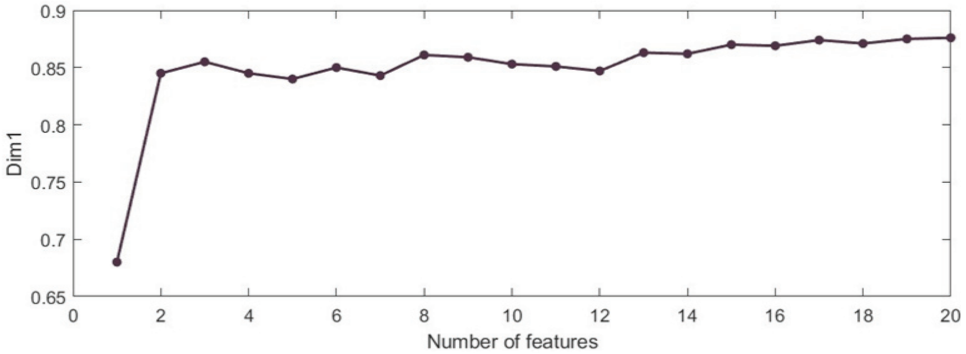

Fig1 Plot of marginal effects of feature importance

3. Empirical analyses

In this study, an innovative approach is adopted where the feature importance scores from the XGBoost algorithm are juxtaposed with those derived from the Random Forest model. Through this comparative analysis, a comprehensive feature selection process is undertaken, amalgamating the strengths of both algorithms. Subsequently, the results are augmented by incorporating the outcomes of the feature forward selection conducted in Step 3. The culmination of these steps culminates in the formulation of a concatenated set, thereby enriching the feature selection process. Finally, the quantitative analysis outcomes are presented in Table 2, depicting the scoring situation table of key items, thus offering a clear insight into the importance and relevance of each feature in influencing the outcomes under study.

Table2 marking scheme

serial number	Feature name	marking scheme
1	LOCALIZER DEV Dots (C)	9.880%
2	LOCALIZER DEV Dots (L)	9.100%
3	Inertial Vertical	8.700%
4	GMT Speed	9.500%
5	LOCALIZER DEV Dots (R)	9.100%
6	PITCH ATT	7.500%
7	ROLL ATT	6.200%
8	N1 SELTED-L	5.500%
9	N1 SELTED-R	3.500%

References

[1] Chen, Lejun, et al. "Flight evaluation of a sliding mode online control allocation scheme for fault tolerant control." Automatica 114 (2020): 108829.

[2] Jung, Ki-Wook, Young-Won Kim, and Chang-Hun Lee. "Aerodynamically controlled missile flight datasets and its applications." International Journal of Aeronautical and Space Sciences 24.1 (2023): 248-260.

[3] Wen, Jiayu, et al. "Hybrid Adaptive Control for Tiltrotor Aircraft Flight Control Law Reconfiguration." Aerospace 10.12 (2023): 1001.

[4] Deng Yang. Research on the group path of influencing factors of general aviation safety accidents[D]. Zhengzhou Institute of Aviation Industry Management, 2024.

[5] Yao Di, Yi Rong, Xiao Yizhuo. Analysis of factors affecting general aviation flight safety[J]. Science and Innovation, 2021(09):23-25.

[6] Yue Qiming, Zhang Quanqing, Zhang Yanwen, et al. Research on the analysis and measurement of factors affecting pilot situational awareness[C]// Aviation Society of China. Proceedings of the Ninth Aviation Society of China Youth Science and Technology Forum.

*__Corresponding author:__ Zihan Zhao(2004.6), female, School of Mathematics and Statistics, Sichuan University of Science & Engineering, research direction: Big Data Statistics.