A Risk Prediction Model of Hard Landing Based on Random Forest Algorithm

Ruishan Sun[1,a], Chongfeng Li[*2,b]

[1]Flight technology collage, Civil Aviation University of China, Tianjin, China
[2]Economics and Management College, Civil Aviation University of China, Tianjin, China

Abstract—Landing safety is a hot issue in civil aviation safety management. In order to fully mine the influence factors of hard landing in flight data and effectively predict the risk of hard landing, the random forest algorithm was introduced. Firstly, this paper qualitatively analyzed the influence factors of hard landing, and chose the features of the model based on the flight data. Secondly, this paper gives a quantitative analysis method of the importance of features based on Gini index. Finally, for the dataset of hard landing was class-imbalanced, the model was training based on SMOTE method. Then, the random forests classifier was built and verified by real flight data. The results showed that the recall rate of the model was 85.50%. The model can not only effectively prevent the occurrence of hard landing, but also provide a method reference for airlines to apply data mining to improve the ability of flight events management.

1 INTRODUCTION

Risk management is the main part of safety management system (SMS) in the civil aviation field, which advocates actively exploring and even predicting the risks in the system to improve the safety management level. Landing safety is the key content of civil aviation safety management. The statistical data of IATA from 2015 to 2019 showed that the accident rate of aircraft in final approach and landing phase was as high as 53%, and the number of hard landing accidents accounted for 23% of the total accidents in landing phrase [1]. According to the servicing manual of Boeing Company, hard landing is a flight event in which the vertical acceleration of an aircraft exceeds the prescribed threshold [2]. It could not only lead to aircraft's structure especially carriage, airfoil and so on in overload, but the human death if severely [3]. Therefore, the prediction of hard landing risk is of great significance to civil aviation safety management.

QAR can record all kinds of performance parameters, environment parameters and operation parameters in the whole flight phase [4]. At present, the diagnosis of hard landing mainly depends on QAR data, especially the vertical acceleration value when the main landing gear touching down to the ground. However, most airlines fail to do the research based on QAR data, which to some extent ignores the risk during the aviate. Civil Aviation Administration of China (CAAC) has implemented the program of Flight Operations Quality Assurance (FOQA) since 1997, with all commercial airplanes of Chinese airlines obliged to install QAR or similar equipment. Practice has proved that QAR data can help airlines improve flight safety management and quality control.

Although many researches on landing safety [5-11] and flight events prediction models [12-15] have been carried out, there are few researches on the analysis of hard landing accidents based on QAR data. For example, Wang et al. [16] pointed out that the vertical load of touching ground was closely related to the touchdown attitude and configuration by logistic model. Liu et al. [17] proposed an improved Bow-tie model, and analyzed the causes and consequences of hard landing by statistical methods, which provided reference for the prevention and risk control of the hard landing. The above research based on traditional mathematical statistics methods inevitably has the disadvantages of poor efficiency and low prediction accuracy of QAR data mining. In recent years, machine learning technology has been gradually applied in the field of civil aviation safety management system to ensure flight safety. Some researchers have studied hard landing using QAR data mining based on machine learning. For example, Chen et al. [18] preprocessed QAR data by correlation analysis and factor analysis, and established SVM model to predict the hard landing. Qiao et al. [19] proposed a new method of hard landing prediction based on RBF neural network with K-means clustering algorithm. Chau et al. [20] developed a deep prediction model based on Long Short-Term Memory (LSTM), and predicted the risk of hard landing based on QAR data.

Random forest is a popular machine learning algorithm, which can be used to develop prediction models [21]. Random forest uses randomly selected training datasets to construct many classification and regression trees, and forecasts by summarizing the results of each tree. Therefore, random forest usually provides higher accuracy compared with the single decision tree model and maintains some useful qualities of the tree model (e.g.,

[a]e-mail: sunrsh@hotmail.com
[b*]Corresponding author: 1049412572@qq.com

ability to interpret relationships between predictors and outcome) [22]. In the setting of classification, random forest always provides higher prediction accuracy compared with other models [23]. One of the main benefits of using random forests for predictive modeling is the ability to handle datasets with many features. In conclusion, the risk prediction of aircraft hard landing based on QAR data is suitable to be regarded as a binary classification problem by using random forest algorithm.

In this paper, the random forest algorithm was applied to the risk prediction of hard landing, and the importance analysis method for the influencing factors of hard landing was given. Meanwhile, the random forest algorithm based on SMOTE method to improve the imbalanced datasets of hard landing were given to realize the risk prediction. This paper could provide a method reference for airlines to improve the level of risk management based on flight data mining.

2 DATA PREPROCESSING

2.1 Selection of features

The premise of risk prediction is to select the features that influence the hard landing from QAR data. Referring to SHEL model [24], the features of hard landing are determined from the perspective of " Human-Equipment-Environment ". Firstly, the human factors mainly include: the control of altitude and speed in the sliding stage; the control column and throttle operation in flare; the flare time and final flare pitch angle. Secondly, for the same aircraft type, the parameters that affect the hard landing are mainly divided into three aspects: attitude, speed, and weight. Finally, the environmental factors that may affect the risk of hard landing include airport's elevation, atmospheric temperature, visibility, wind shear, icing and so on.

Based on the above analysis, the ground speed (*GS*), vertical speed(*IVV*), pitch angle(*PITCH*), pitch rate(*PITCHR*), roll angle(*ROLL*), flaps angle(*FLAP*), total air temperature(*TEM*), static air pressure(*PRE*), longitudinal wind speed(*LW*), glide deviation(*GD*), gross weight(*GW*), control column position (*CP*) and throttle column position (*TP*) were selected as the features of the model.

2.2 QAR Data Collection and Preprocessing

The 128 cases of QAR data in this study were collected from the Boeing 737-800 fleet of a local airline. The original data is a CSV (Comma Separated Value) file with thousands of rows and columns. Therefore, Python programing functions in Microsoft Excel was applied. According to the FOQA [25], the threshold of determining hard landing for this aircraft type was set as 1.6 g in this study. Landing samples was extracted in 128 cases at 1 Hz below 50 feet and we totally got 1401 samples including 23 hard landing samples. In principle, random forest algorithm is not sensitive to the unit and dimension of data, so it does not need to normalize the sorted data.

3 RISK PREDICTION METHOD OF HARD LANDING BASED ON RANDOM FOREST ALGORITHM

3.1 The Basic of Random Forest

Random forests classifier (RFC) uses the Bootstrap method to form different training sets, train decision trees respectively and vote to form the result. RFC algorithm steps are as follows：

(1) Generate decision trees

1. The Bootstrap method is used to randomly select k samples from the datasets to form the training set $\boldsymbol{T}$;

2. Output each group of data in the training set with tree structure:

a. m attribute features are randomly selected from n attribute features;

b. The Gini index of m attribute features is used as the standard for the best node of single tree;

c. Repeat the previous step to divide the dataset into binary trees until the tree nodes are pure.

(2) The set $\boldsymbol{H}$ formed by k decision trees is called the random forest.

(3) The classification results of all decision trees in the random forest are counted, and the voting result of the trees is the final classification result of the random forest.

3.2 Risk Prediction Model of Hard landing Based on Random Forest Algorithm

3.2.1 Processing method imbalance datasets

The existing research seldom considers the problem of imbalanced datasets for hard landing samples in real QAR data, which will directly affect the prediction accuracy of machine learning algorithms such as random forest. SMOTE (synthetic minority oversampling technique) is an oversampling technique for synthesizing minority classes. It is an improved scheme based on the random oversampling algorithm. Because the random oversampling adopts the strategy of simply copying samples to increase a small number of samples, it is easy to be over-fitting. The basic idea of SMOTE algorithm is to analyze the minority samples and add new samples to the datasets according to the minority samples. The algorithm flow of expanding the hard landing samples is as follows:

(1) For the hard landing sample x, the distance from Euclidean distance to all samples in a few sample sets is calculated, and its k-nearest neighbor is obtained.

(2) Set a sampling scale according to the imbalanced proportion of the hard landing sample to determine the sampling ratio n. For each small sample x, select several samples randomly from its k neighbor, assuming that the nearest neighbor selected is x_n.

(3) For each randomly selected neighbor x_n, the new samples are constructed according to formula (1) with the original sample respectively:

$$x_{new} = x + rand(0,1) \times (x_n - x) \quad (1)$$

3.2.2 Importance analysis of features

In order to quantitatively describe the importance of hard landing features, the average change of Gini index is used to measure the feature importance for RFC. The larger Gini index is, the higher the feature importance is. *VIM* is used to represent variable importance measures, and *GI* is Gini coefficient. Suppose there are *c* features $X_1, X_2, \ldots, X_c$. Now we need to calculate each feature's Gini index score ($VIM_j^{(GI)}$), that is, the average change of node branching purity in all decision trees of RFC. The calculation method of Gini index is shown in formula (2):

$$GI(\mathrm{m}) = 1 - \sum_{k=1}^{|K|} p_{\mathrm{m}k}^2 \tag{2}$$

Where *K* is the number of categories and $P_{\mathrm{m}k}$ is the proportion of category *k* in node m.

The importance of feature X_j in node m, that is, the variation of Gini index before and after node m branching, is shown in formula (3):

$$VIM_{j\mathrm{m}}^{(GI)} = GI_{\mathrm{m}} - GI_{\mathrm{l}} - GI_{\mathrm{r}} \tag{3}$$

Where, GI_{l} and GI_{r} is the new Gini index after node m branching. If the nodes of feature X_j appearing in decision tree i are in set M, then the importance of X_j in the tree i is shown in formula (4):

$$VIM_{ij}^{(GI)} = \sum_{\mathrm{m}\in M} VIM_{j\mathrm{m}}^{(GI)} \tag{4}$$

Suppose that there are *n* trees in RFC, the calculation method of $VIM_j^{(GI)}$ is shown in formula (5):

$$VIM_j^{(GI)} = \sum_{i=1}^{n} VIM_{ij}^{(GI)} \tag{5}$$

Finally, the normalized results of all the obtained importance scores are shown in formula (5):

$$VIM_j = \frac{VIM_j^{(GI)}}{\sum_{i=1}^{c} VIM_i} \tag{6}$$

So far, the feature importance ranking of *c* features can be obtained to assist airlines to control the risk of hard landing purposefully.

3.2.3 Calculation of model parameters

Random forest algorithm needs to determine the optimal combination of two parameters, namely, the number of decision trees(n_{tree}) in random forest and the number of features (n_{try}) for branching from all the features in the process of generating the subtree of random forest. At the same time, Breiman research shows that random forest can often get better results under the default parameter setting[21].

The parameter n_{tree} is the number of trees in the random forest. The generalization error of the random forest will converge to an upper bound with the increase of n_{tree}, so that the random forest will not over fit. Therefore, the goal is to choose a large enough n_{tree} to make the OOB error tend to be stable. The n_{tree} should not be too large, too many trees will increase the training time of the model. First, n_{try} is set as the square root of the number of features, then n_{tree} of different groups is selected to establish a random forest model. Finally the trend of OOB error of each group is observed to determine the value of n_{tree}.

The parameter n_{try} is the number of features selected randomly each time when the tree branches. Random selection of features makes the difference between trees larger, and improves the noise tolerance and generalization ability of the model, so the selection of this parameter is very important. Based on the determination of n_{tree}, the optimal n_{try} is selected to minimize the OOB error by enumeration.

3.2.4 Training and verification of the model

Based on the balanced QAR datasets improved by SMOTE, the optimal n_{tree} and n_{try} are obtained, the RFC is established. Breiman has proved that the accuracy of using OOB error to verify the model is the same as using the test datasets [26]. The prediction accuracy *P*, recall *R* and comprehensive evaluation index F_1 of RFC are calculated to verify the availability of the model. The calculation formula of *P*, *R*, and F_1 is shown in formula (7):

$$\begin{cases} P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{P} \\ F_1 = \dfrac{2PR}{P + R} \end{cases} \tag{7}$$

Where *TP* indicates that the positive class is predicted as positive class, *FP* means that the negative class is predicted as positive class. *P* is the proportion of the real hard landing samples divided into hard landing samples; *R* means the classification accuracy rate of the actual hard landing samples; F_1 is the comprehensive evaluation index of the model, and the higher *P, R* and F_1, the more effective RFC is.

4 EXPERIMENTAL ANALYSIS

Based on the balanced datasets obtained by SMOTE method, the proportion of normal landing and hard landing samples was 52.39% and 47.60% respectively.

The importance of features calculated according to the equation (6) is shown in Figure 1:

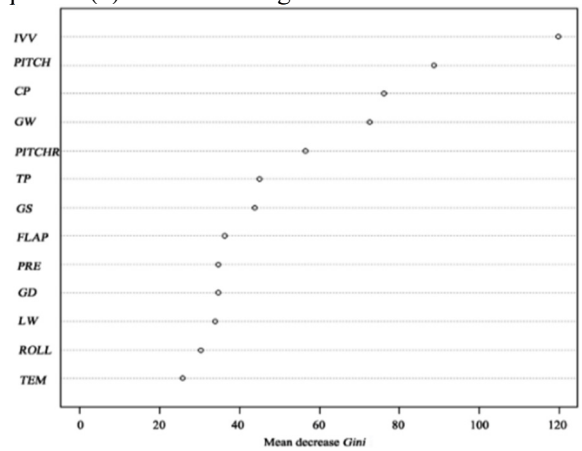

Fig 1. The importance of features in risk prediction model of hard landing

The results show that the vertical speed, pitch angle, control column position, gross weight and pitch rate have great influence on the hard landing risk. Airlines can control the risk from the above five aspects and reduce it to an acceptable level such as improving flight training methods.

The preset parameter n_{try} of the model was the square root of the number of features, that is, n_{try}= 3. The RFC was established by setting n_{tree} of different groups. The OOB error of each group is shown in Figure 2. When n_{tree} = 500, the OOB error of the model tends to be stable. On this basis, the change of model accuracy under different n_{try} values is shown in Figure 3. When n_{try} = 3, the model OOB error is the smallest. So, the optimal parameters of the model are n_{tree} = 500, n_{try} = 3.

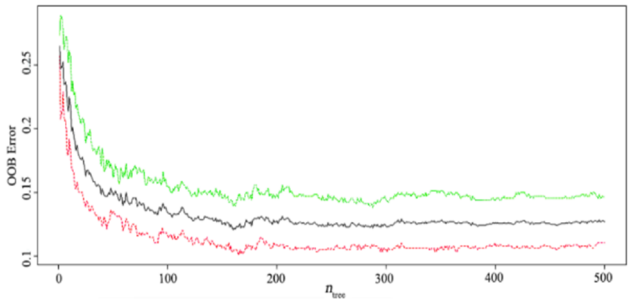

Fig 2. The relationship between n_{tree} and OOB error

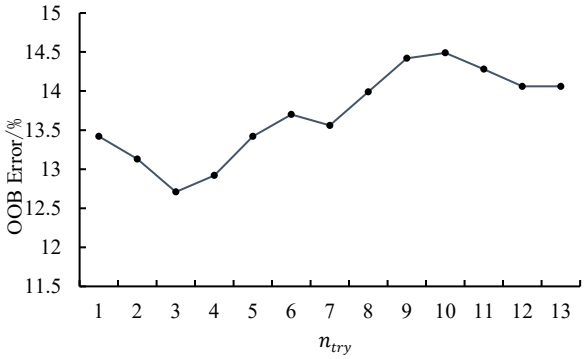

Fig 3. The relationship between n_{try} and OOB error

After the model training, the OOB error of the model decreased to 12.71%. Compared the predicted results with the real results and draw a cross table as shown in Table 1:

Table1. Cross table of predict results and real results

Actual Value \ Prediction	Normal landing	Hard landing	Error rate/%
Normal landing	653	81	11.0
Hard landing	97	570	14.5

At the same time, table 2 compares the performance of the model between the original datasets and the improved datasets based on SMOTE method. We can see that the model has a higher *R* based on the improved datasets.

Table2. Evaluation index of the results in different training datasets

Datasets	*R*/%	*P*/%	F_1/%
Original	26.09	98.78	41.27
Improved	85.50	87.06	86.27

Taking the real landing process of a certain flight as an example, the prediction results of the hard landing risk category are illustrated in table 3. The hard landing risk is marked as "high", and the normal landing risk is marked as "low". After 50 feet the hard landing risk was "low"; in the 6th second, it turned to "high" and the warning lasted for two seconds until the aircraft touched the ground. Compared with the real results, the risk prediction of hard landing of the flight is misjudged in the 7th second, and the overall prediction effect of the model is acceptable.

Table3. The result of the hard landing risk prediction model in one flight

Time/s	*GS*	*IVV*	*PITCH*	*PITCHR*	*ROLL*	*FLAP*	*TEM*	*PRE*	*LW*	*GD*	*GW*	*CP*	*TP*	Prediction value	Actual value
Unit	KNOT	FEET/MIN	DEG	DEG/SEC	DEG	DEG	°C	MB	KNOT	DOT	LBS	DEG	DEG	/	/
1	146.500	-576.000	0.879	0.703	-1.230	240.117	31.500	1002.394	1.977	-0.589	113360.000	1.798	47.637	Low	Low
2	146.500	-464.000	1.934	0.234	-0.879	240.117	31.500	1002.903	2.473	-0.938	113360.000	1.776	45.000	Low	Low
3	145.500	-320.000	2.285	1.172	-0.352	240.117	31.500	1003.339	2.977	-1.107	113360.000	3.013	39.551	Low	Low
4	144.500	-256.000	2.813	-0.352	0.703	240.117	31.500	1003.666	3.989	-1.661	113360.000	4.103	36.387	Low	Low
5	143.500	-224.000	2.285	0.410	1.055	240.117	31.500	1003.848	4.997	-1.393	113360.000	5.372	36.387	Low	Low
6	142.000	-64.000	2.988	0.469	1.055	240.117	31.750	1003.994	5.500	-2.223	113360.000	4.880	36.387	High	High
7	140.500	-112.000	2.285	-0.762	0.176	240.117	31.750	1003.884	6.000	-1.571	113360.000	3.930	36.387	High	Low
8	GROUND														

5 DISCUSSION AND CONCLUSION

In order to predict the risk of hard landing scientifically, a prediction model based on random forest classifier was constructed, including the selection method of features, the improvement method of imbalanced datasets, the importance analysis method of features, and the construction method of the model. Based on the real QAR data, the training and validation of the model were completed. The results showed that the recall rate of the model was 85.50%, which had a acceptable ability to

predict the risk of hard landing.

The method of constructing risk prediction model of flight events based on random forest can acquire the importance of features, and has the advantages of simple implementation, high accuracy, and strong comprehensive performance. It can be extended to other risk prediction work based on flight data mining, to continuously improve the level of civil aviation risk management.

This study did not fully consider the impact of pilot's operating characteristics. In the future, we can mine the human factors affecting hard landing based on QAR data so that we can further improve the accuracy of the model. At the same time, we can develop QAR flight event risk analysis system based on big data and continuously strengthen the civil aviation safety management ability based on data-driven.

ACKNOWLEDGMENT

The flight data for this paper are supported by the safety management department of a local airline. We couldn't disclose its name out of confidentiality, but we are grateful for its help.

REFERENCES

1. IATA. Safety Report 2019. International Air Transport Association: Montreal-Geneva;2020.
2. Wang, L. . (2017). Effects of flare operation on landing safety: a study based on Anova of real flight data. Safety Science, 102(2018).
3. Hu, C. , Zhou, S. H. , Xie, Y. , & Chang, W. B. . (2016). The study on hard landing prediction model with optimized parameter SVM method. Control Conference. IEEE.
4. Wang, L. , Wu, C. , & Sun, R. . (2014). An analysis of flight quick access recorder (qar) data and its applications in preventing landing incidents. Reliability Engineering & System Safety, 127, 86-96.
5. C.E. Dole, J.E. Lewis Jr, J.R. Badick, B.A. Johnson, Flight Theory and Aerodynamics: A Practical Guide for Operational Safety, John Wiley & Sons,2016.
6. L. Witte. R. Roll, J. Biele, S. Ulamec, E. Jurado, Rosetta lander Philae-Landing performance and touchdown safety assessment, Acta Astronaut.125 (2016)149-160.
7. N. Trawny, A. Huertas, M.E. Luna, C.Y. Villalpando, K.E. Martin, J.M. Carson, A.E. Johnson, C. Restrepo, V.E. Roback, Flight testing a real-time hazard detection system for safe lunar landing on the rocket-powered morpheus vehicle, in: Proc. of AlAA Science and Technology Forum and Exposition,2015.
8. L. Li, R. Hansman, R. Palacios, R. Welsch, Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring, Transp. Res.C64(2016)45-57.
9. Y. Dai, J. Tian, An analysis method of landing safety based on rough set theory.in: Reliability and Maintainability Symposium, RAMS, 2012 Proceedings - Annual, IEEE, 2012. pp.1-6.
10. L. Yi, S. Zhang, L Xueqing, A hazard analysis-based approach to improve the landing safety of a BWB remotely piloted vehicle, Chin. J. Aeronaut. 25(6)(2012)846-853.
11. T. Patterson, S. McClean, P. Morrow, G. Parr, Modelling safe landing zone detection options to assist in safety critical UAV decision making. Procedia Comput. Sci.10(2012)1146-1151.
12. G. Holmes, P. Sartor, S. Reed, P. Southern, K. Worden, E. Cross, Prediction of landing gear loads using machine learning techniques, Struct. Health Monit.15(5)(2016)568-582.
13. E. Cross, P. Sartor, K. Worden, P. Southern, Prediction of landing gear loads from flight test data using Gaussian process regression, in: Proceedings of the International Workshop in Structural Health Monitoring,2013.
14. F. Herrema, V. Treve, R. Curran, H. Visser, Evaluation of feasible machine learn-ing techniques for predicting the time to fly and aircraft speed profile on final approach, in: International Conference for Research in Air Transportation,2016.
15. A. Nanduri, L. Sherry, Anomaly detection in aircraft data using Recurrent Neural Networks(RNN), in: Integrated Communications Navigation and Surveil-lance, ICNS,2016, IEEE,2016, pp.5C2-1.
16. Lei, W. , Wu, C. , Sun, R. , & Cui, Z. . (2014). An Analysis of Hard Landing Incidents Based on Flight QAR Data. International Conference on Engineering Psychology and Cognitive Ergonomics. Springer, Cham.
17. Liu J. , Zhang L. . Analysis of civil aviation hard landing events based on improved bow tie model. Journal of Transport Information and Safety, 2016, 34 (4): 57-62
18. Hu, C. , Zhou, S. H. , Xie, Y. , & Chang, W. B. . (2016). The study on hard landing prediction model with optimized parameter SVM method. Control Conference. IEEE.
19. Qiao, X. , Chang, W. , Zhou, S. , & Lu, X. . (2016). A prediction model of hard landing based on RBF neural network with K-means clustering algorithm. 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE.
20. Chao, T. , Xiang, Y. , Li, J. , Zhu, T. , Lv, R. , & Liang, S. , et al. (2018). An innovative deep architecture for aircraft hard landing prediction based on time-series sensor data. Applied Soft Computing, 73, S1568494618304654-.
21. Breiman, L. . (2001). Mach. learn. Machine Learning, 45, 5-32.
22. Speiser, J. L. , Durkalski, V. L. , & Lee, W. M. . (2015). Random forest classification of etiologies for an orphan disease. Statistics in Medicine, 34(5).
23. Fernandez-Delgado, M. , Cernadas, E. , Barro, S. , &

Amorim, D. . (2014). Do we need hundreds of classifiers to solve real world classification problems?. Journal of Machine Learning Research, 15, 3133-3181.

24. Molloy, G. J. , & O??Boyle, C. A. . (2005). The shel model: a useful tool for analyzing and teaching the contribution of human factors to medical error. Academic Medicine Journal of the Association of American Medical Colleges, 80(2), 152-5.
25. Ac-121/135-fs-2012-45r1. Implementation and management of Flight Operational Quality Assurance (FOQA)[S]. 2015.6.30: 20.
26. Fang, K.N. , Wu, J.B. , Zhu, J.P. . Review of random forest method. Statistics & Information Forum, 2011, 26(003):32-38.