Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.34716

Prediction of Space Mission Success Rate Based on Random Forest

Mingze Tang

Tianjin Fayao High School, Tianjin, China
shdjnptmz@gmail.com

Abstract. Space exploration has received much attention globally, with numerous nations investing heavily in the advancement of the space industry. Safety is one of the important aspects in any space mission that calls for keen investigation. Accurately predicting the success rate directly affects task planning, resource allocation, and risk management. Regarding the current situation, there are few predictions about the success rate of space missions, leaving a significant gap. In this study, the Random Forest algorithm is used to analyze the Kaggle Global Space Exploration Dataset (2000-2025). The data set contains 12 attributes and 3000 observations, which can be used in predicting the success rate of space missions. Using visual techniques, the effectiveness of the Random Forest algorithm in minimizing the safety risk related to space missions is assessed. The results show that the Random Forest algorithm can be applied in predicting the success rate of space missions. It is worth noting that although the Random Forest algorithm performs well, its stability poses a challenge.

Keywords: Random Forest, space exploration, Kaggle

1. Introduction

In the light of the development of technology and modern times, the industry of aeronautics has made significant improvements in the field of technical sciences. Nevertheless, the high cost and high risks involved in the process of researching and launching spacecraft remain the problems to be addressed in future. Numerous incidents involving the loss of control and explosion of spacecraft have taken countless human lives. Each of those accidents is worth in-depth analysis. That is why an attempt to predict the possibility of mission success based on the indicators of previous space missions would be an efficient solution to the problem. In this respect, one should consider the application of Random Forest algorithm to make predictions. For instance, in 2025, Mingqiang C; Wenhao Z; Yanjun S; Haodong L; Zhonghang D applied the Random Forest algorithm to predict the fuel consumption of Cessna 172 general aviation aircraft, proving its effectiveness [1]. Also, in 2025, Yang S and Rongze Z applied the Random Forest algorithm to assess fire risk for the aircraft on the basis of some indicators, which helped identify five main factors affecting the safety of the operation [2]. Bergmayr T, Höll S, Kralovec C and Schagerl M used the Random Forest algorithm to assess and predict strain residuals, allowing to check the structure and making sure it was safe for further use in operating large-scale spacecraft [3]. Also, in 2020, Dong X, Taylor C J, and Cootes T

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.34716

F. applied the Random Forest algorithm to detect any defects in linear thin welds and reduce the workload for aerospace engineers [4]. In 2019, Dasari S K, Cheddad A, and Andersson, P applied the Random Forest algorithm to develop surrogate models in the framework of aerospace industry using the datasets from the Turbine Rear Structure use case [5]. At the same time, the number of research works on the safety of spacecraft is rather low. That is why the current research paper concentrates on aerospace research and attempts to predict the success rate of space missions using the Random Forest algorithm.

2. Methodology

2.1. Data description

In this case, the dataset used has been obtained from the highly esteemed Kaggle database which is one of the popular databases that contains data from different industries. In total, as shown in Table 1 there are 12 features in this dataset, with 3000 observations contained within the dataset. This data belongs to the domain of space exploration around the globe in the period ranging from 2000 to 2025 [6].

Table 1. Dataset variables

Variable name	Type	Description
Country	Categorical	Nation involved in the space mission
Year	Numerical	Year of the space mission (2000-2025)
Mission name	Identifier	Name of the mission
Mission type	Categorical	Manned / Unmanned
Launch site	Categorical	Location of launch
Satellite type	Categorical	Communication / Navigation / Research
Budget (in Billion $)	Numerical	Mission budget
Success rate (%)	Numerical	Mission success percentage
Technology used	Categorical	Reusable Rocket / AI Navigation / Solar Propulsion
Environmental impact	Categorical	Low / Medium / High
Collaborating countries	Categorical	
Duration (in Days)	Numerical	

2.2. Feature engineering

The engineering of features is an important part of any machine learning algorithm that contributes significantly to its efficiency. In this regard, it means choosing the appropriate features from the available data in order to increase the performance of the algorithm [7]. As far as foundational machine learning algorithms go, there is a great emphasis on ensuring that the features contribute positively towards the effectiveness of the algorithm. First of all, one should start with the preprocessing of redundant variables in order to enhance the speed of training the model. Among the 12 variables analyzed, the variable named Mission name is redundant since it is not helpful and may even negatively affect the algorithm. In addition, the problem of missing values needs to be addressed. Therefore, the first step is to conduct a preliminary check to detect missing values. Upon completion of the analysis, it became evident that there are no missing values in the data. Hence, the

DOI: 10.54254/2755-2721/2026.34716

task becomes easier when performing the feature engineering. In the paper, the author attempted to measure the impact of international cooperation on the success rate of missions using the data provided in the 'Collaborating Countries' column. To begin with, one needed to establish whether some major spacefaring countries took part in the missions. After completing the task, the first column, namely, the 'Collaborating Countries', was eliminated.

After the conclusion of data preprocessing, the subsequent stage includes the process of training on the data. The data contains 3000 instances and can be termed as relatively small in terms of size. To ensure the ease of training, the data set is partitioned into 80 percent training data and 20 percent testing data. In effect, there exist 2400 instances for training and 600 instances for testing.

2.3. Random Forest

Random Forest is an example of an ensemble method. It consists of several decision trees, and the accuracy of the output values is achieved through the voting or averaging technique.

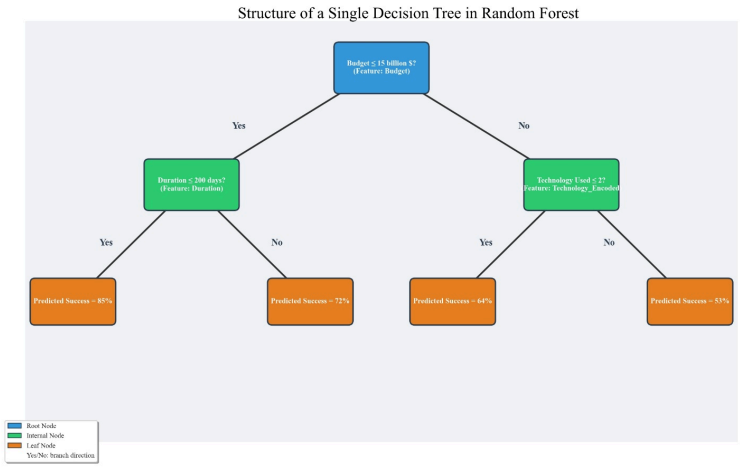


Figure 1. Structure of a single decision tree in Random Forest

In Figure 1 a single decision tree in a Random Forest is used as an example to explain what nodes are. In a decision tree, each node branches in two ways that are designated YES or NO. The first decision that takes place at the root node is if the budget is less than or equal to $1.5 billion. If the answer to this decision is yes, then the tree moves forward to the internal node where a similar question concerning if the task duration is less than or equal to 200 days is asked. Such questions continue being asked along the decision tree up to when one reaches the leaf node. The final outcome will be obtained through summing up the success values from the leaf nodes. Several decision trees make up a Random Forest algorithm. Random Forest algorithm is usually preferred to a regression algorithm due to its better performance and efficiency [8].

In this experiment, the candidate values considered for the number of trees include 100 and 200. As far as the candidate values for the tree maximum depth are concerned, those considered in this experiment include 10, 20, and None. Furthermore, the candidate values chosen for the minimum number of samples needed for splitting an internal node included 2, 5, and 10. Finally, the top 10 candidate prediction results were evaluated by comparing them with the actual results. In order to assess the accuracy of the Random Forest algorithm, negative mean squared error criterion was used.

DOI: 10.54254/2755-2721/2026.34716

2.4. Performance evaluation

An in-depth analysis is performed in this paper regarding the effectiveness of the Random Forest method for predicting the success rate of space missions. The analysis process will employ the use of four visualization methods and comparison charts.

Firstly, in order to demonstrate the effectiveness of the prediction through the Random Forest model, ten test cases based on the success rate were selected. The predicted success rates and actual success rates of these ten test cases were then compared against each other. The absolute errors and differences of each test case were derived.

$$e_i = \left| y_i - \hat{y}_i \right| \tag{1}$$

$$d_i = \hat{y}_i - y_i \tag{2}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| \tag{3}$$

Also, another efficient method in identifying the important factors affecting the success rate of the task is through the feature importance analysis using the 12 variables involved. To avoid the occurrence of Hughes effect causing excessive training time and inefficiency in training process [9, 10]. In this paper, the study is concentrated on finding the top 10 features with high score values. The outcome of the experiment will be briefly discussed using tabulation below. Application of the technique on the Feature Importance plot. Make use of color in order to improve visualization. Determine the top 10 features with the highest score values. Assign different colors to three feature types in Table 2.

Table 2. Correspondence between colors and feature types in the feature importance plot

Color	Feature type
Bule	Numerical
Orange	Categorical
Green	Engineered

The next step towards investigating the abilities and accuracies of prediction for the Random Forest model on the success rates of space mission ventures is done by using the scatter plot graph. The scatter plot graph is one of the visualization methods that is popularly used due to its ability to illustrate relations and patterns among variables. As seen in the graph below, the predicted success rate and actual success rate are plotted along y-axis and x-axis, respectively. The reference line was utilized as means for illustrating perfect predictions, enabling comparison based on prediction error. Moreover, a linear fit line was used as another method for comparing the data visually. The 95% confidence interval band indicates the accuracy of predictions [11]. In addition, MAE and relative error measures can be applied as other statistical measures to assess prediction performance.

DOI: 10.54254/2755-2721/2026.34716

$$\hat{y} = a + bx \tag{4}$$

$$\hat{y}_0 \pm t_{0.025,n-2} \times SE_{\hat{y}0} \tag{5}$$

$$Relative\ Error = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \times 100\% \tag{6}$$

Additionally, residual analysis was done to determine the level of compliance with the assumptions of the Random Forest model. The process involved the plotting of a Q-Q graph whereby the residuals were placed in an ascending manner, and then the values were plotted along the x-axis. The points in the graph were represented using the blue color while the red line assisted in interpreting the plot by serving as a benchmark. The closer the points were to the red line, the nearer their conformity to the normality assumption.

$$t_i = \Phi^{-1}\left(\frac{i-0.5}{n}\right) \tag{7}$$

Finally, a residual scatter plot was generated to analyze how many of the data values lay above and below the zero value. This plot also facilitated identifying any outlier data values lying away from the zero line.

3. Experimental results and analysis

3.1. Prediction results for the first 10 test samples

Table 3 shows the accuracy rates for the top ten samples, which represent the comparison between the real accuracy rate and the estimated one. It is clear that there are certain differences between the actual values and the predicted ones, reflecting the ability of the algorithm to make accurate predictions about the actual success rates. The greatest difference found in this analysis is -23.03, whereas the least difference is -5.61, suggesting that there were some differences in the predictions. There were some absolute errors made during the prediction process, ranging from 5.61 to 23.03, giving an MAE of 13.35.

Table 3. Actual vs predicted success rate (first 10 test samples)

Variable No.	Actual success rate(%)	Predicted success rate(%)	Difference	Absolute Error
1	80.00	74.39	-5.61	5.61
2	64.00	75.68	+11.68	11.68
3	58.00	78.89	+20.89	20.89
4	100.00	76.97	-23.03	23.03
5	58.00	75.39	+17.39	17.39
6	82.00	72.48	-9.52	9.52

DOI: 10.54254/2755-2721/2026.34716

Table 3. (continued)

7	62.00	73.93	+11.93	11.93
8	80.00	73.73	-6.27	6.27
9	86.00	78.16	-7.84	7.84
10	94.00	74.72	-19.28	19.28
Mean Absolute Error(MAE)				13.35

3.2. Feature importance analysis

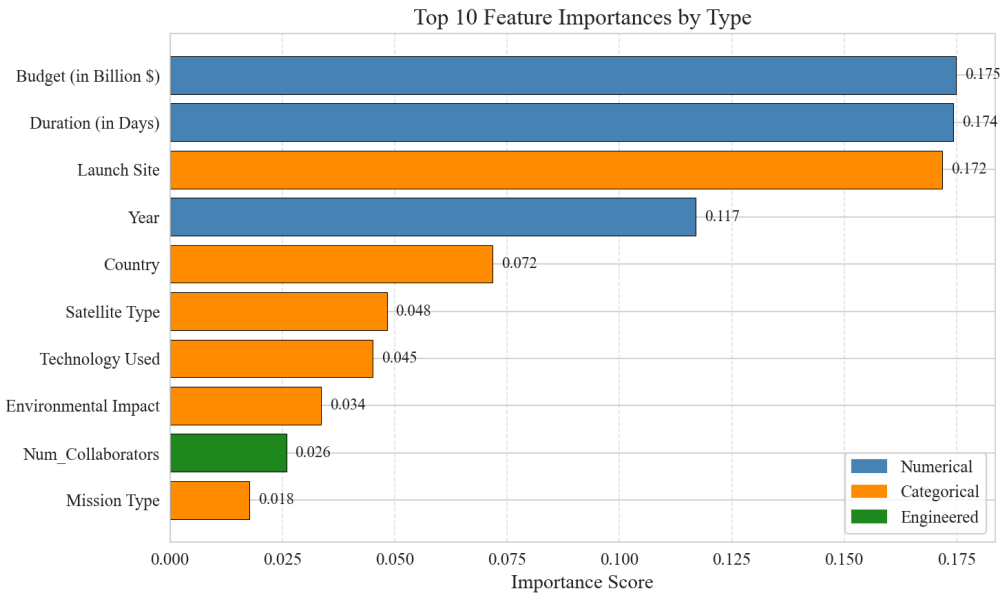


Figure 2. The importance scores of the top 10 features

Figure 2 gives the complete illustration for the top ten features and their scores respectively. As shown in the above figure, there exists a considerable difference between the scores attained by the top three features and those obtained by the other remaining features. This suggests that the top three features are very important compared to other features. The first feature which comes in the list of top features is 'Budget (in Billion$)', whose score is 0.175, making it the most important feature. On the other hand, 'Duration (in Days)' has a score of 0.174 and hence stands as the second most important feature, while 'Launch Site' has a score of 0.172, making it the third most important feature. On the contrary, the 'Mission Type' has a lower score of 0.018, thereby rendering it as one of the least important features and should not be included during further analysis since it is unnecessary. Moreover, it can be observed from the above figures that numerical features get the highest score when compared to categorical features. This implies the importance of taking numerical features into consideration while analyzing data.

DOI: 10.54254/2755-2721/2026.34716

3.3. Prediction scatter plot

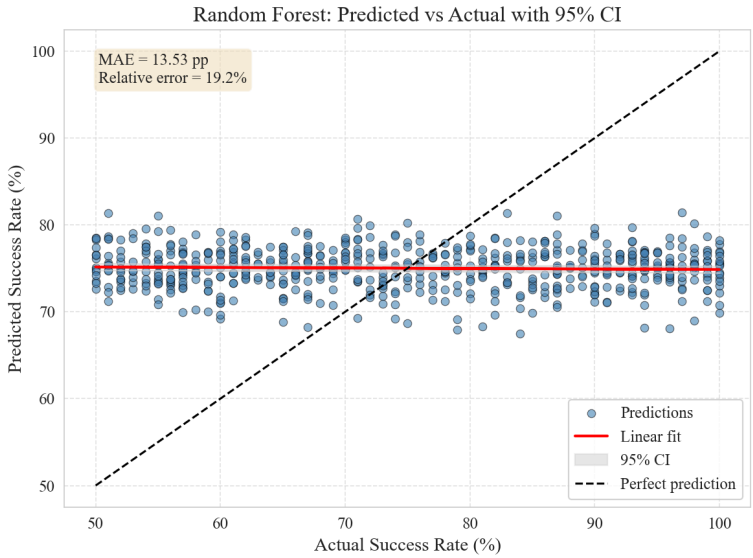


Figure 3. Prediction scatter plot in Random Forest

Figure 3 shows the predicted values, which are observed to have a parallel distribution. Both the linear fit line and 95% confidence interval have parallel distributions in relation to the prediction, and the gradient is equal to zero. It can be inferred from the above result that the Random Forest model does not appropriately account for the challenges in the success of space missions.

3.4. Residual analysis

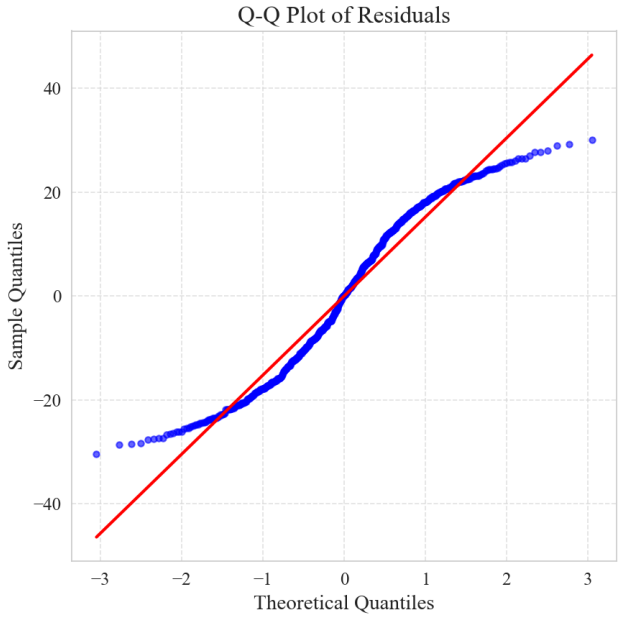


Figure 4. Q-Q plot in Random Forest

DOI: 10.54254/2755-2721/2026.34716

Figure 4 below can be used to see the relationship between the normal quantiles that are theoretically computed on the x-axis and the sample quantiles measured on the y-axis. The blue scatter dots are used to indicate the distribution of the residuals and the red dots indicate that the residuals are almost normally distributed. In the current case, the red line deviates insignificantly in a spiral way in relation to the blue line.

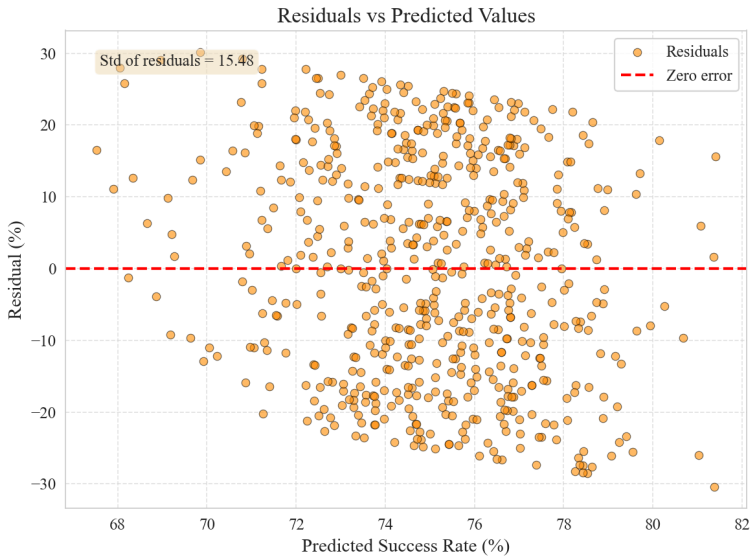


Figure 5. Residual scatter plot in Random Forest

Figure 5 shows the Residual on the y-axis, which is the difference between the Residual and the estimated success rate. Conversely, the x-axis is the prediction of the success rate by the model. The value of the red line is zero, which means that the actual values are exactly equal to the predicted values. Though overall scattered, there are cases when the residuals do not have a significant difference with the zero line.

4. Conclusion

This paper has employed Random Forest algorithm to estimate the success rate of space missions. The study begins by discussing the workings of the Random Forest algorithm and then a description of a simple decision tree and how a Random Forest works. The success of the Random Forest algorithm usage is also proved by the comparison of 10 pairs of predictions with the actual values in this study. The average error value of this process is 13.35 which means that the Random Forest algorithm can be used to identify the success rate of space missions.

Then, four visualizations are created to aid the analysis of the data generated. Firstly, there is the feature importance score chart, in which the top 10 most important features based on the scores obtained are displayed. These features have been evaluated by critically analyzing their scores and types. It is noteworthy that the feature that requires in-depth analysis is named Budget (in Billion$). In addition, a prediction scatter plot is created to check the successfulness of the Random Forest model in the prediction of space mission success rate. A problem fitting has been witnessed here, which should be further enhanced. In addition, a residual analysis is also performed with a Q-Q Plot of the residuals and a residual scatter plot. The reference line showed a clear deviation on the Q-Q plot; this implies that the success of the space missions can be better predicted with the help of the Random Forest model, in terms of accuracy.

DOI: 10.54254/2755-2721/2026.34716

Finally, although the model had been successful in estimating the success rates of space missions, it has its weaknesses, such as instability. These weaknesses can be explained by the fact that the data set has issues, and it is important to enhance the data to achieve the optimal prediction results.

References

[1] Mingqiang, C., Wenhao, Z., Yanjun, S., Haodong, L., & Duanzhong, H. (2025). Fuel consumption prediction for general aviation aircraft based on regression trees and Random Forest-dndc. Science Technology & Engineering, 25(7).

[2] Yang, S., & Rongze, Z. (2025). Comprehensive Evaluation of Aircraft Fire Risk Based on Random Forest Algorithm. Science Technology & Engineering, 25(13).

[3] Bergmayr, T., Höll, S., Kralovec, C., & Schagerl, M. (2023). Local residual Random Forest classifier for strain-based damage detection and localization in aerospace sandwich structures. Composite Structures, 304, 116331.

[4] Dong, X., Taylor, C. J., & Cootes, T. F. (2020). A Random Forest-based automatic inspection system for aerospace welds in X-ray images. IEEE Transactions on Automation Science and Engineering, 18(4), 2128-2141.

[5] Dasari, S. K., Cheddad, A., & Andersson, P. (2019, May). Random Forest surrogate models to support design space exploration in aerospace use-case. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 532-544). Cham: Springer International Publishing.

[6] Kaggle, Global Space Exploration Dataset (2000-2025). https: //www.kaggle.com/datasets/atharvasoundankar/global-space-exploration-dataset-2000-2025?resource=download

[7] Verdonck, T., Baesens, B., Óskarsdóttir, M., & vanden Broucke, S. (2024). Special issue on feature engineering editorial. Machine learning, 113(7), 3917-3928.

[8] Segal, M., & Xiao, Y. (2011). Multivariate Random Forests. Wiley interdisciplinary reviews: Data mining and knowledge discovery, 1(1), 80-87.

[9] Shahshahani, B. M., & Landgrebe, D. A. (2002). The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Transactions on Geoscience and remote sensing, 32(5), 1087-1095.

[10] Tadjudin, S., & Landgrebe, D. A. (2002). Covariance estimation with limited training samples. IEEE Transactions on Geoscience and Remote Sensing, 37(4), 2113-2118.

[11] Sachs, M. C., Brand, A., & Gabriel, E. E. (2022). Confidence bands in survival analysis. British Journal of Cancer, 127(9), 1636-1641.