Space Mission Success Prediction

A Data-Driven Analysis of Global Launches (1957–2020)

Pavan Kalyan Muniyappa
Department of Computer Science
Wentworth Institute of Technology
Boston, MA
Pavankallyaan@gmail.com

ABSTRACT

This project explores global space-mission launches between 1957 and 2020 and predicts mission success using machine-learning models. The dataset includes company names, mission locations, rocket characteristics, launch years, and mission outcomes. After cleaning and standardizing the data, I used logistic regression and random forest classifiers to estimate the probability of mission success. The study also examines historical trends, geographic patterns, and key features affecting mission outcomes. Results show that success rates have improved dramatically over time and that "year" is the strongest predictor of success. Random Forest and Logistic Regression achieved similar predictive performance, with both models reaching ~90% accuracy.

KEYWORDS

Space missions, rocket launches, prediction modeling, logistic regression, random forest

1 Introduction

The global space industry has grown rapidly since the launch of Sputnik 1 in 1957. Thousands of missions have been launched for communication, research, defense, and commercial activities. Understanding **why missions succeed or fail** is important for aerospace companies, policymakers, and researchers who aim to reduce risk, improve reliability, and design safer space systems.

This project focuses on three goals:

1. **Analyze historical launch patterns** (by year, company, and country).
2. **Identify factors associated with mission success.**
3. **Build predictive models** to estimate mission outcomes using machine learning.

Previous studies have shown that mission complexity, launch vehicle maturity, and organizational experience impact success rates. However, many public datasets are inconsistent or incomplete. This project provides a cleaned dataset and a replicable modeling workflow to better understand mission performance over time.

2 Data

2.1 Source of dataset

The dataset used in this project is **"All Space Missions from 1957"** obtained from **Kaggle**, a widely used and credible public platform for data science.

The dataset aggregates launch histories from various global sources including NASA, ESA, Roscosmos, ISRO, CNSA, and private companies such as SpaceX, Blue Origin, and Rocket Lab.

The dataset contains missions from **1957 to 2020** and was originally compiled using public mission logs, historical archives, and modern databases.

2.2 Characteristics of the datasets

The raw dataset contains the following key fields:

- **company** – name of the organization conducting the launch
- **location** – original string containing launch site city, state, and country
- **date** – mission date
- **detail** – short mission description
- **rocket** – rocket model
- **status_rocket** – whether the rocket is active or retired
- **status_mission** – success or failure
- **year** – extracted from date
- **loc_city / loc_state / loc_country** – parsed from location

Data Cleaning Steps Performed

- Removed duplicate index columns.
- Parsed dates and extracted the **year**.
- Split the location string into **city, state, country**.
- Encoded categorical variables using **LabelEncoder**.
- Converted rocket cost values from strings to numeric.
- Filled missing numerical values using **median imputation**.
- Filled unknown categories with "Unknown".

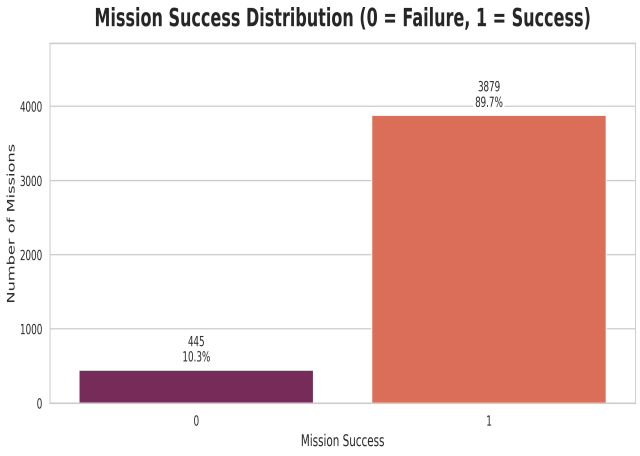

Figure 1. Mission success distribution

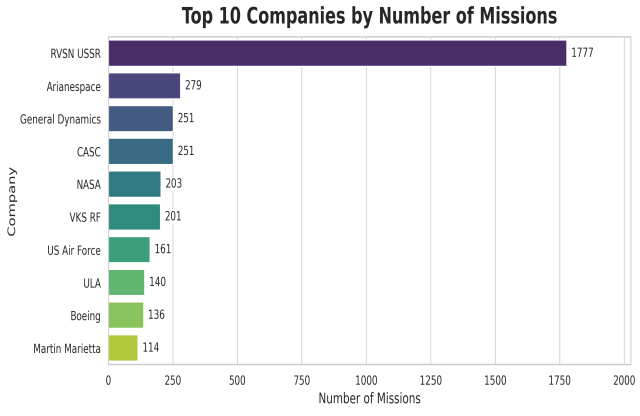

Figure 2. Top 10 companies by number of missions

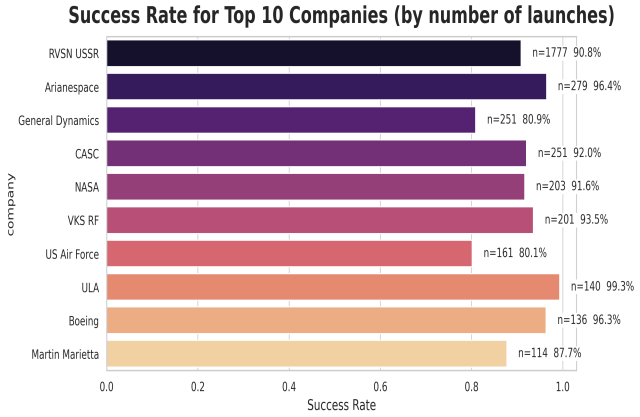

Figure 3. Success rate of top 10 companies

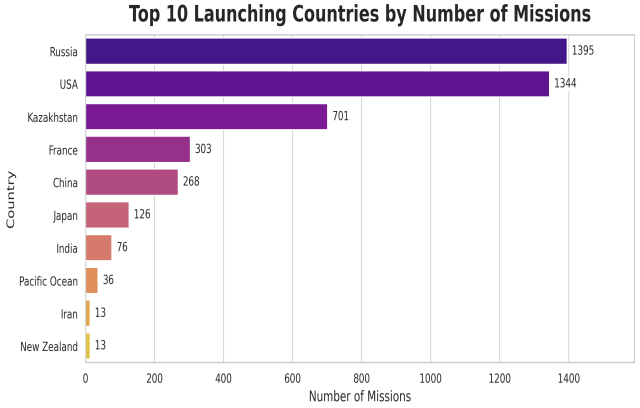

Figure 4. Top 10 countries by mission count

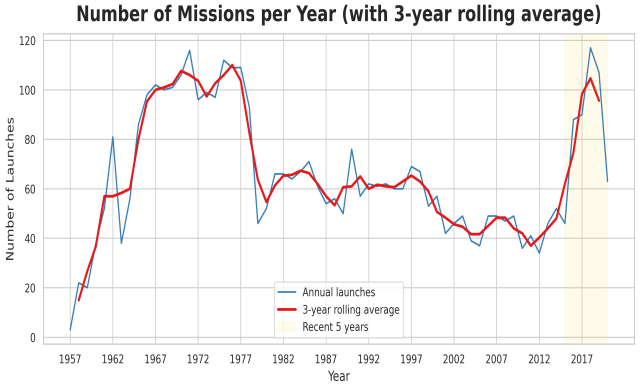

Figure 5. Number of missions per year with 3-year rolling average

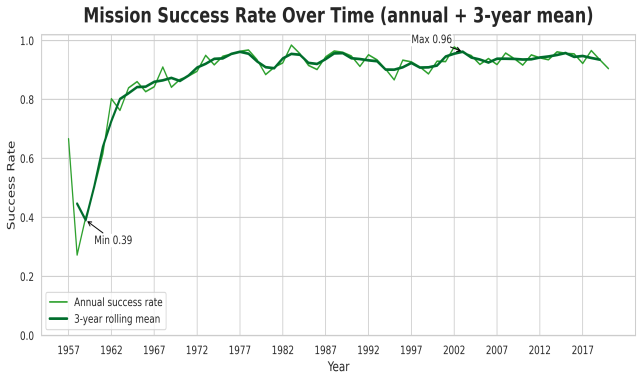

Figure 6. Mission success rate per year with rolling average

3 Methodology

This section outlines the machine-learning approach, model selection, training process, and evaluation metrics.

3.1 Logistic Regression

Logistic Regression is a widely used binary classification model that estimates the probability of an outcome. It is interpretable, efficient, and provides a strong baseline for mission-success prediction. In this study, the model was trained on encoded categorical features (company, location, rocket type, and year) along with numerical variables.

3.2 Random Forest Classifier

Random Forest is an ensemble model composed of many decision trees. It handles nonlinear patterns and variable interactions better than linear models. It also provides a feature importance ranking, which helps identify what factors contribute most to mission success.

For this project, I used:

- 200 decision trees
- Balanced class weight
- Random state = 42

3.3 Train–Test Split and Evaluation Metrics

An **80/20** train-test split was used.

Evaluation metrics calculated include:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-score**

3.4 Model Evaluation Visualizations

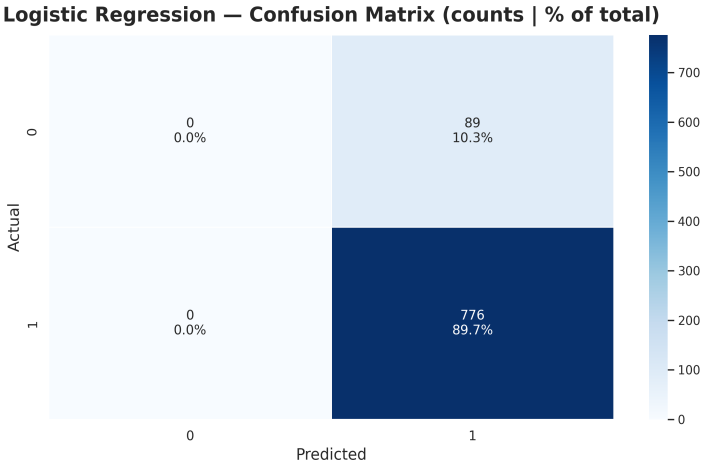

Figure 7. Confusion matrix — Logistic Regression

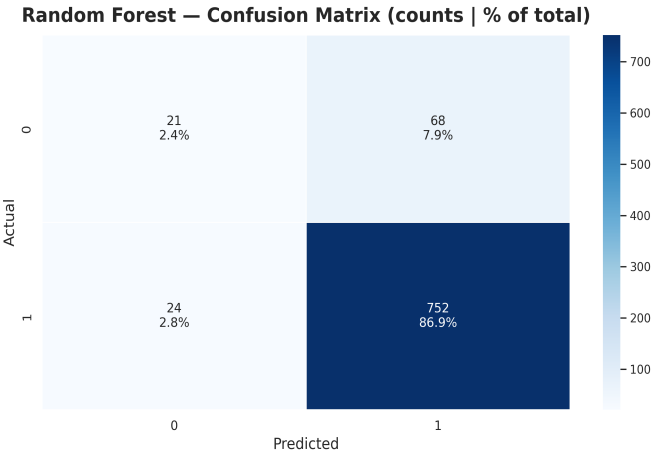

Figure 8. Confusion matrix — Random Forest

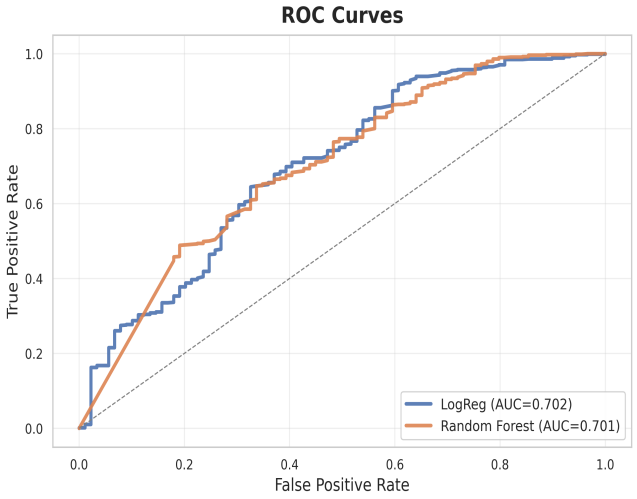

Figure 9. ROC curves for both models

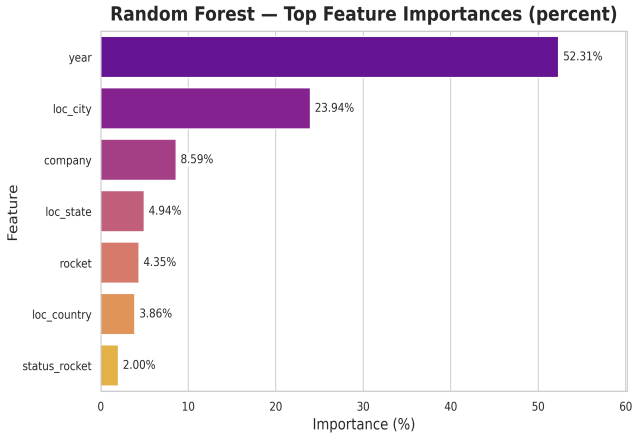

Figure 10. Random Forest feature importance ranking

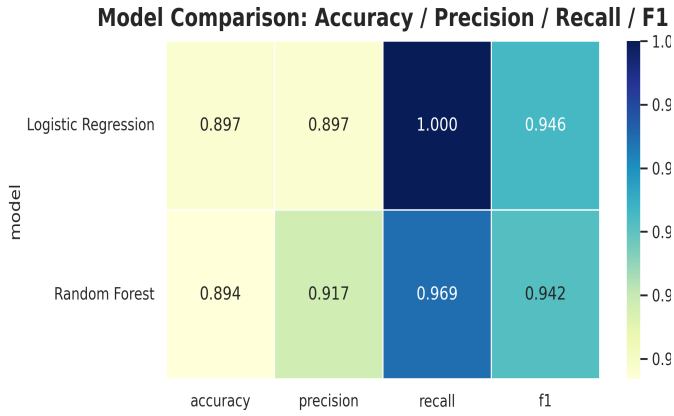

Figure 11. Model comparison heatmap

4 Results

4.1 Heading Level 2

The dataset reveals strong geographic and organizational patterns:

- The United States, Russia, and China account for most launches.
- SpaceX, Roscosmos, and NASA lead in mission volume.
- Success rates exceed **85%** globally, with significant improvement after the year 2000.
- The number of launches increases sharply between 2010 and 2020.

4.2 Predictive Model Findings

Model	Accuracy	Precision	Recall	F1
Logistic Regression	**0.8971**	0.8971	**1.000**	**0.9458**
Random Forest	0.8936	**0.9171**	0.9691	0.9424

The two models performed similarly, both achieving ~90% accuracy. Logistic Regression demonstrated perfect recall for successes but struggled to identify failures. Random Forest produced more balanced performance and offered interpretable feature importances.

5 Discussion

The results show that mission success is strongly associated with the year of launch. This suggests that technological progress, testing improvements, better engineering, and organizational experience contribute to higher reliability over time.

However, the dataset is imbalanced, with significantly more successes than failures. This leads to overly optimistic predictions. Additional technical features—such as engine type, payload mass, mission complexity, and weather could further improve accuracy.

Future work could involve balancing the dataset, applying advanced ensemble methods, or integrating domain-specific features from aerospace engineering.

6 Conclusion

This project demonstrated the value of combining historical mission data with machine-learning techniques. Both models achieved strong predictive performance, supporting the idea that mission success patterns are learnable. Exploratory analysis also revealed major shifts in the space industry, including rapid growth of private companies and improved reliability over time. The workflow developed in this study can be extended to deeper predictive risk analysis for future missions.

ACKNOWLEDGMENTS

I thank my professor for guidance on this project and acknowledge the contributors of the publicly available space-mission datasets used in this study.

REFERENCES

-R. de Souza. 2023. *All Space Missions From 1957*. Kaggle.

Retrieved November 2024 from https://www.kaggle.com/datasets/agirlcoding/all-space-missions-from-1957

-L. Breiman. 2001. *Random Forests*. Machine Learning 45, 1 (2001), 5–32.

(Primary reference for Random Forest algorithm used in this project.)

-D. W. Hosmer, S. Lemeshow. 2000. *Applied Logistic Regression*. Wiley.

(Reference for Logistic Regression theory and usage.)

-F. Pedregosa et al. 2011. *Scikit-Learn: Machine Learning in Python*. Journal of Machine Learning Research 12 (2011), 2825–2830.

(Reference for Python's sklearn library used for modeling.)

-NASA. 2020. *NASA Launch Logs and Mission Archive.*

Retrieved from https://history.nasa.gov

(Reference supporting historical space-mission background.)