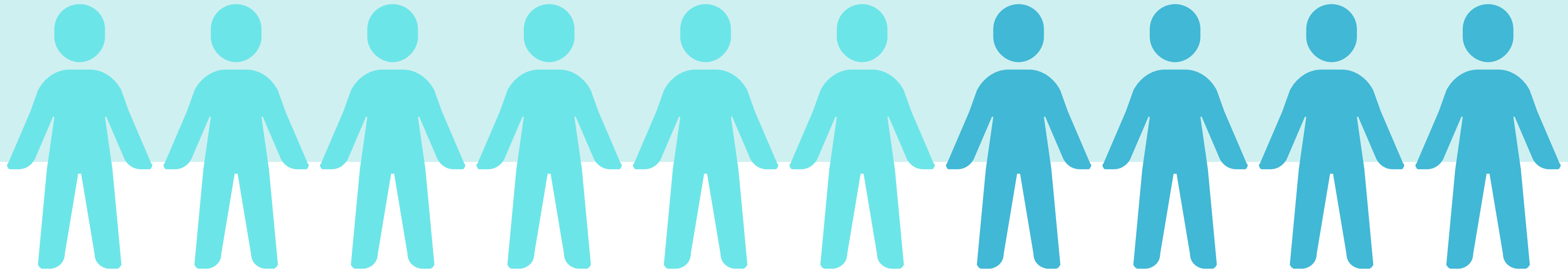
GLOBAL SPACE EXPLORATION SUCCESS RATE PREDICTION

By Danish Najmuddin

4/1/2026

PROBLEM STATEMENT

“Predicting the probability of success for a space mission is critical because failures can cause massive financial losses, delays, and reputational damage, affecting space agencies, private companies, investors, insurers, satellite operators, and mission planners, and without accurate predictions, resources may be wasted and critical missions jeopardized.”



DATA OVERVIEW

Source: https://www.kaggle.com/datasets/atharvasoundankar/global-space-exploration-dataset-2000-2025/data (Kaggle)

- Granularity: Each row represents a single space mission, including details about the rocket, satellite, launch company, and other mission-specific attributes.
- Size: 3,000 rows × 12 columns
- Target Variable: Mission success probability

OBJECTIVES AND KEY QUESTIONS

Project Objectives:

1. Predict the probability of success for space missions based on rocket specifications, satellite details, and launch company.
2. See which factors are linked to success.

Key Analytical Question:

1. Which rocket, satellite, and launch company features are most strongly associated with mission success?

METHODOLOGY

Preprocessing & EDA (default)
EDA Focus:

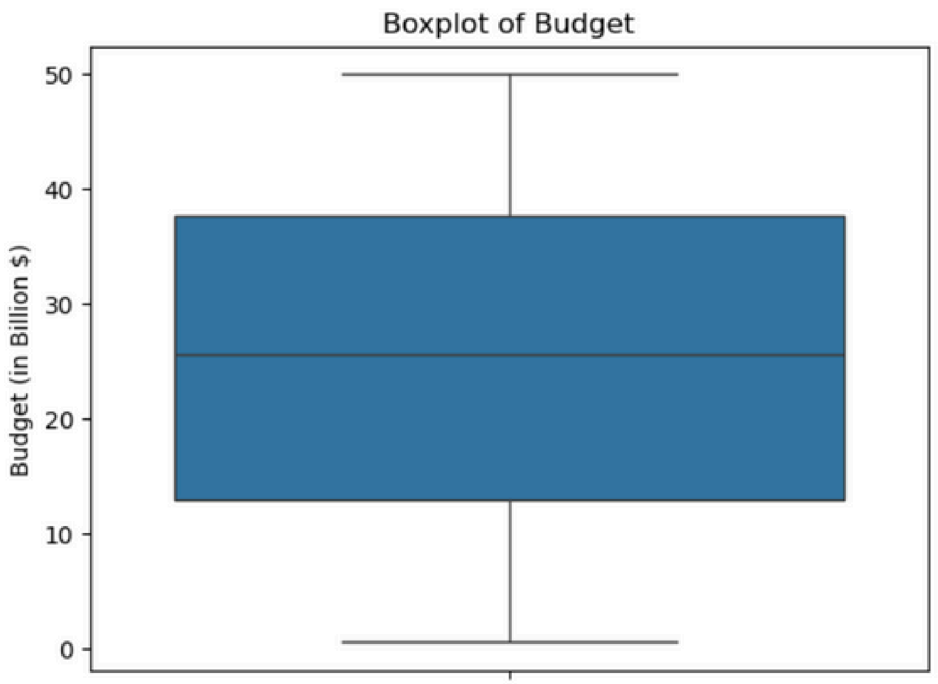
- Plot class distribution of mission based on probability of success
- Time-series of success rate by year.
- Bar charts of success rate by country and by mission type (satellite, crewed, research).
- Correlation / pairplots for numeric features like budget and mission success percentage

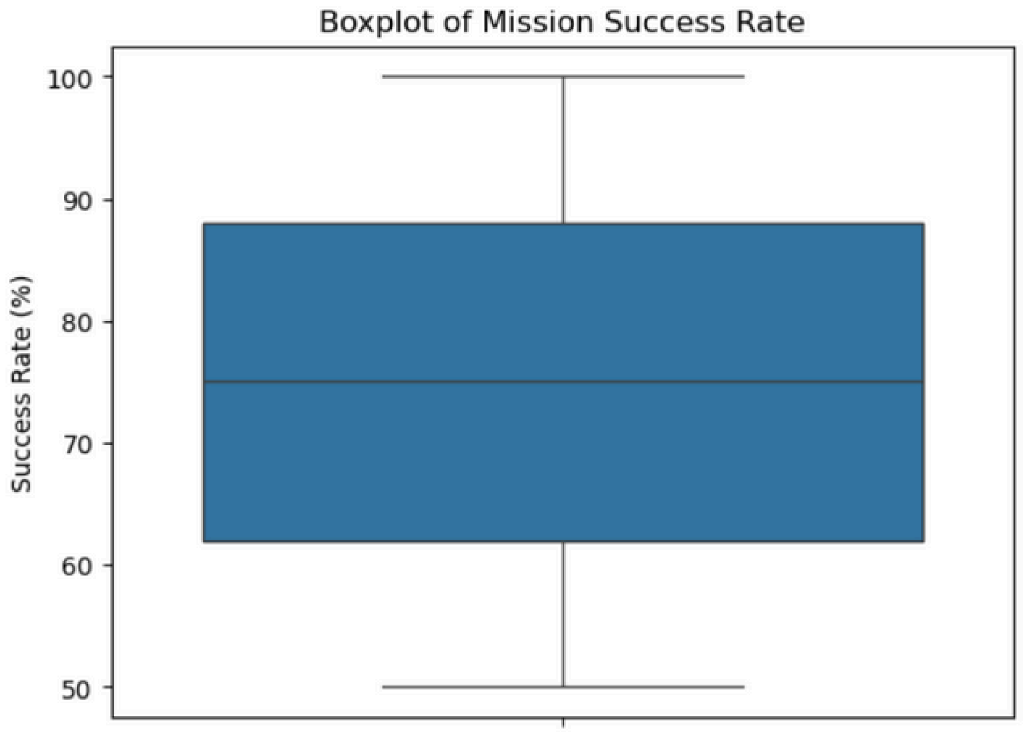
METHODOLOGY

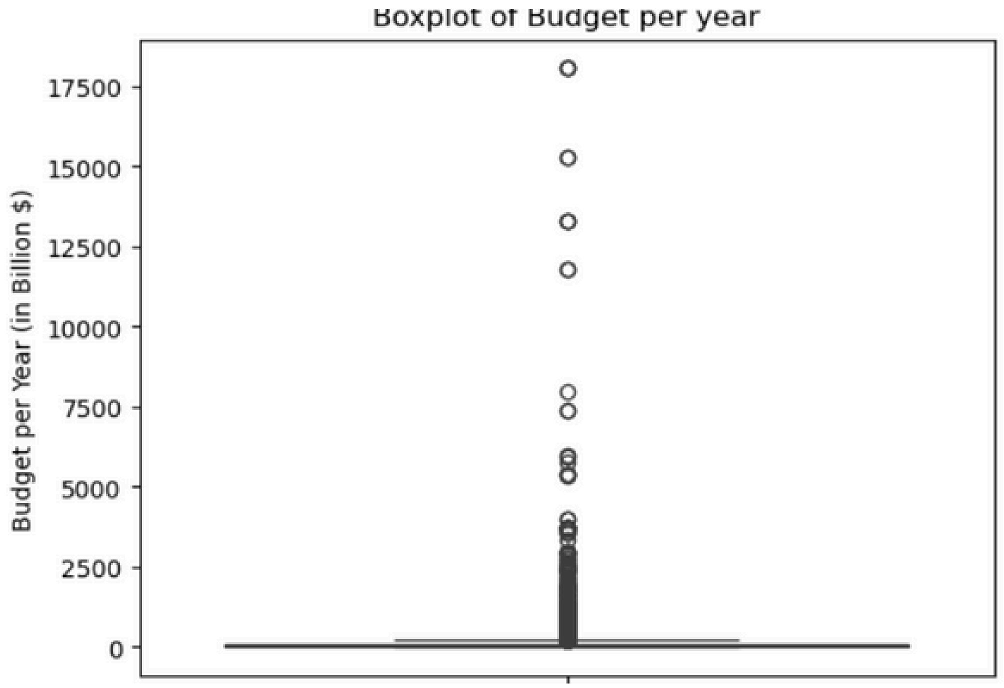
Feature Engineering:
•Clean missing values (e.g., fill / drop for budget and tech fields).
•Encode categorical columns (country, mission type, tech used like reusable rocket/AI nav).
•Create new columns such as:
- “Budget per year” (budget divided by mission year range).
- “High_budget_flag” (low/medium/high turned into numeric).

Deployment:
- Application: I will be using Gradio for the application.
- User Interaction: User selects or types mission info (country, missiontype, budget level, tech like reusable rocket). The app predicts the probability of success.

EDA KEY FINDINGS







EDA KEY FINDINGS

These plots tell us that the original values (success rate, original budget) have no major outliers, while the newly added values (budget PER YEAR) has a lot of major outliers.

This tells me that I don't really have too much to worry about, since I am only using my newly added values for EDA, and I will only use the original values for machine learning

EDA KEY FINDINGS

```
df.isna().sum()
```

```
Country                    0
Year                       0
Mission Name               0
Mission Type               0
Launch Site                0
Satellite Type             0
Budget (in Billion $)      0
Success Rate (%)           0
Technology Used            0
Environmental Impact       0
Collaborating Countries    0
Duration (in Days)         0
dtype: int64
```

```
df.duplicated().sum()
```

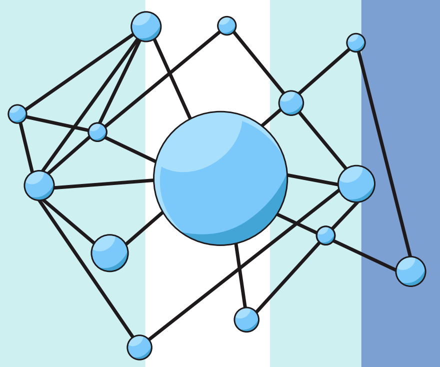
```
0
```

There are no duplicated or missing values from the dataset.

We can conclude that this is a very high quality dataset.

MODELING APPROACH

- Algorithms: Tried Linear Regression as a baseline, then Random Forest and GridSearchCV to capture non-linear relationships and improve performance.
- Validation: Used an 80/20 train–test split to evaluate generalization on unseen mission data.
- Feature Engineering: Applied One-Hot Encoding for nominal features, Ordinal Encoding for ordered categories, and removed filler columns to reduce noise.



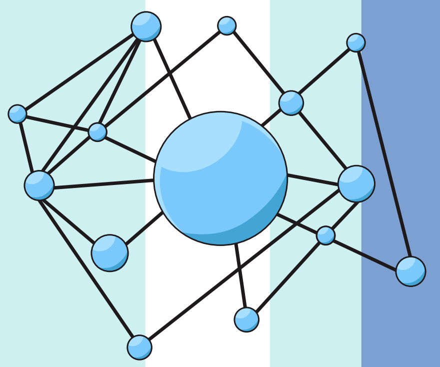
RESULTS & EVALUATION

- Primary Metrics (Regression):
  - MAE: 8.44
  - RMSE: 10.58
  - R^2: 0.51
- Model Plot: Feature importance plot showing budget, duration, and technology as key drivers.
- So What Insight: The model predicts mission success rate within ~±10%, helping decision-makers estimate mission risk before launch.


Satellite Mission Success Prediction App
Main Country
UK
Year
0
Mission Type
Unmanned
Satellite Type
Research
Budget (in Billion $)
0
Technology Used
Traditional Rocket
Environmental Impact
Low
Collaborating Country
Germany
USA
Russia
Japan
UAE
Israel
UK
India
China
France
Predicted Success Rate (%)
0
Flag

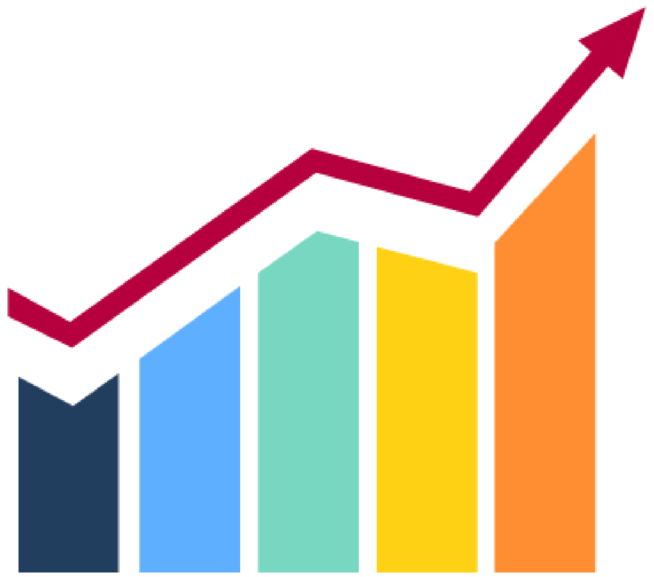
MEASURE OF SUCCESS

- Target Metric: Achieve RMSE ≈ 10 or lower on unseen data.
- Result: Achieved RMSE = 10.58 and R^2 = 0.51, meeting acceptable predictive performance for a real-world, noisy dataset.
- Business KPI: Enables early evaluation of mission configurations to identify high-risk missions before committing budget and resources.



CHALLENGES & LIMITATIONS

- Encoding multiple categorical features – solved with OneHotEncoder and OrdinalEncoder.
- Ensuring unseen user inputs don’t break the model – used handle_unknown='ignore' in encoders and used set drop down inputs instead
- Struggled with making the app - learnt and understood how to make the Gradio app properly



FUTURE WORK & RECOMMENDATIONS

If I had more time to improve this project, I would make the app using Streamlit instead of Gradio and make the app a lot more interactive

TECH STACK

- Programming Language: Python
- Libraries:
  - Data Manipulation & Analysis: Pandas, NumPy
  - Visualization: Matplotlib, Seaborn, Plotly
  - Machine Learning: Scikit-Learn, XGBoost, LightGBM
  - Model Evaluation: Scikit-Learn metrics (Accuracy, F1 Score, AUC-ROC)
- App / Deployment: Gradio for live demo
- Infrastructure / Tools: Git for version control, Jupyter Notebook for development, Kaggle for dataset hosting

Q&A

THANK YOU!