**SpaceX First Stage Landing Prediction Report**

**By**

**Muhammad Ibthesam**



# Abstract

The reuse of rocket boosters marks a paradigm shift in the economics of space exploration, reducing launch costs and enabling more frequent missions. SpaceX, through its Falcon 9 program, has been a global leader in reusable launch vehicle technology. A critical component of this innovation is the successful landing of the Falcon 9 first stage. Predicting the likelihood of landing success provides both operational benefits and a valuable case study in applied data science.

This project presents a comprehensive pipeline that integrates multiple stages of data analysis: **data collection** via APIs and web scraping, **data wrangling** to clean and prepare raw information, **exploratory data analysis (EDA)** to uncover trends, **interactive visual analytics** to engage stakeholders, and **machine learning modelling** to predict outcomes. Several classification algorithms, including Logistic Regression, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbours (KNN), were tested to assess their performance.

The study highlights key variables influencing booster landing success, including payload mass, orbit type, launch site, and booster reuse history. Results demonstrate that machine learning models can provide accurate predictions and valuable insights into engineering decisions. In addition, interactive dashboards and geographic maps serve as practical tools for communicating findings to both technical and non-technical audiences.

Table of Contents

[Abstract 2](#_Toc207480167)

[1. Introduction 4](#_Toc207480168)

[2. Methodology 4](#_Toc207480169)

[2.1 Data Collection 4](#_Toc207480170)

[2.2 Data Wrangling 5](#_Toc207480171)

[2.3 Exploratory Data Analysis (EDA) 5](#_Toc207480172)

[2.4 Visual Analytics 6](#_Toc207480173)

[2.5 Machine Learning Modeling 6](#_Toc207480174)

[3. Results 7](#_Toc207480175)

[4. Discussion & Insights 7](#_Toc207480176)

[5. Conclusion 8](#_Toc207480177)

# 1. Introduction

The high costs of spaceflight have historically been driven by the expendability of rocket hardware. Traditionally, each launch required brand-new boosters, resulting in inefficiencies and costs running into hundreds of millions of dollars. SpaceX disrupted this model by developing reusable rockets, where the Falcon 9 first stage booster is designed to return safely after launch and be prepared for reuse.

Accurately predicting the outcome of these landings has both practical and scientific value. From a business perspective, it allows better planning and risk management. From a research perspective, it demonstrates the application of **data-driven decision-making** in a highly technical field.

This project integrates concepts from computer science, data engineering, and machine learning to address the problem of predicting Falcon 9 landing success. Unlike traditional aerospace engineering approaches, this project leverages **open data and machine learning models** to draw insights.

# 2. Methodology

## 2.1 Data Collection

The foundation of this study was a rich dataset compiled from multiple sources:

* **SpaceX REST API**: Provided structured JSON records for launches, including payload mass, orbit, mission outcomes, and landing attempts.
* **Web Scraping (Wikipedia)**: Using Beautiful Soup, historical Falcon 9 launch data was scraped to supplement missing records and validate API data.
* **CSV Files and SQL Databases**: Provided structured tables with payload details, site codes, and mission outcomes. SQL queries were executed to merge these records efficiently.

Together, these sources produced a dataset containing details for over 100 Falcon 9 launches, each annotated with mission, payload, and landing outcome attributes.

## 2.2 Data Wrangling

Raw data was incomplete and inconsistent. Wrangling steps included:

* Handling **missing values** (e.g., unknown outcomes were imputed or excluded).
* **Standardizing categories**, such as orbit names and landing site codes.
* **Feature engineering**, e.g., creating binary columns like *Landing Success* (1 = successful landing, 0 = failure).
* **Merging datasets**, ensuring that each launch had consistent information across payload, orbit, site, and outcome.
* Applying SQL queries to filter, aggregate, and validate structured data.

This produced a **clean master dataset** that was ready for exploration and modelling.

## 2.3 Exploratory Data Analysis (EDA)

EDA sought to answer guiding questions:

* Which launch sites had the highest landing success rates?
* How does payload mass affect the probability of success?
* Does orbit type influence landing reliability?
* Are there correlations between reuse history and outcomes?

Visualizations included:

* **Bar charts** of landing success rates by site.
* **Scatter plots** of payload mass vs. landing success.
* **Heatmaps** showing correlations between variables.
* **Histograms** displaying distributions of payload and mission success.

A graph of orange and blue dots

AI-generated content may be incorrect.

Figure : Different Success Rates

A graph of blue bars

AI-generated content may be incorrect.

Figure : Success Rate of each orbit

A graph showing the growth of a stock market

AI-generated content may be incorrect.

Figure : Success Rate over the years

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure : Features

## 2.4 Visual Analytics

Interactive tools were used to enhance insights:

* **Folium Maps** displayed geographic distributions of launch sites, allowing spatial reasoning about why ocean-based drone landings are more difficult.
* **Dash Dashboard** allowed stakeholders to interactively filter data by launch site, orbit, or year, and instantly view success probabilities.

These tools created a **data-driven storytelling experience**, transforming raw datasets into **actionable insights** for both technical engineers and decision-makers.

A pie chart with different colored circles

AI-generated content may be incorrect.

A screen shot of a graph

AI-generated content may be incorrect.

Figure : Dashboard

## 2.5 Machine Learning Modelling

The predictive modelling process included:

* **Feature Engineering**: Variables included payload mass, orbit, site, booster reuse, and number of previous flights.
* **Algorithms Tested**:
  + Logistic Regression – baseline interpretability.
  + Decision Trees – rule-based classification.
  + Support Vector Machines (SVM) – strong predictive power with tuned kernels.
  + K-Nearest Neighbours (KNN) – distance-based classification.
* **Model Evaluation**: Confusion matrices, ROC curves, and accuracy metrics were generated for each model.
* **Cross-Validation** ensured models were generalizable and not overfitted.
* **Hyperparameter Tuning** improved performance, e.g., optimizing tree depth or SVM kernel type.

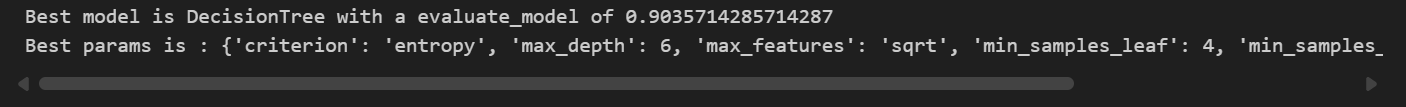


Figure : Best Model

Note: That the score may differ whenever they are trained and can give different result.

# 3. Results

The machine learning models demonstrated varying degrees of predictive success:

* Logistic Regression achieved moderate accuracy but provided interpretability.
* Decision Trees highlighted clear rules (e.g., heavy payloads in GTO orbits had lower success rates).
* SVM achieved the **highest accuracy** after hyperparameter tuning, outperforming other classifiers.
* KNN was sensitive to parameter *k* but performed reasonably well.

Key predictors of success included **payload mass, orbit type, and launch site**.

A screenshot of a black screen

AI-generated content may be incorrect.

Figure : Accuracies

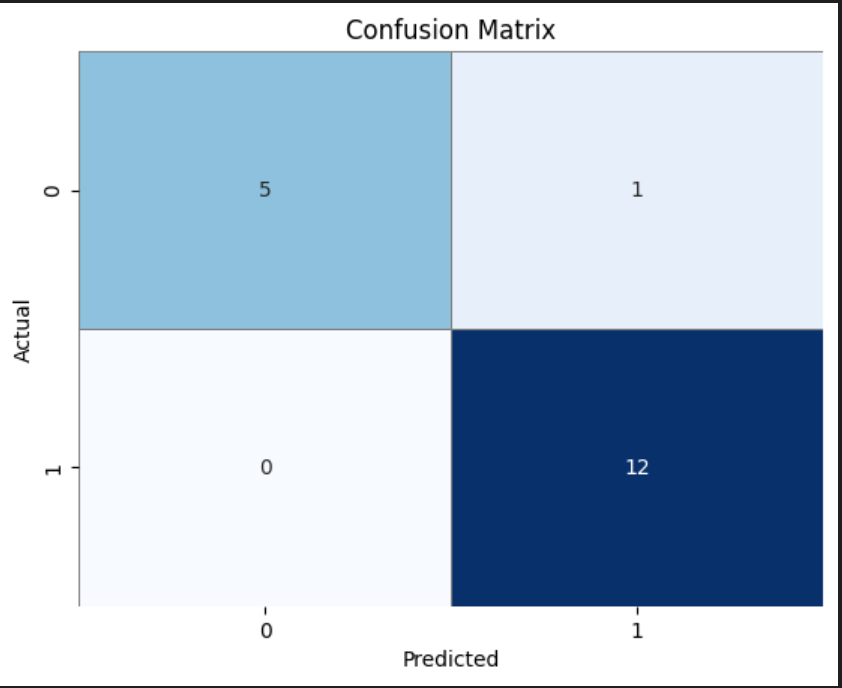


Figure : Confusion Matrix of best score

# 4. Discussion & Insights

Several insights were gained:

* **Launch Site Infrastructure**: Coastal drone ship landings had lower success rates due to weather and sea-state challenges.
* **Payload Weight**: Heavier payloads reduced fuel margins, leading to fewer successful returns.
* **Orbit Differences**: Missions to LEO had significantly higher landing success compared to GTO or interplanetary trajectories.
* **Model Trade-offs**: Simpler models offered interpretability, while SVM provided stronger performance.

This combination of **EDA, visual analytics, and predictive modelling** offered a robust understanding of Falcon 9 landings and highlighted the broader role of machine learning in aerospace.

# 5. Conclusion

This project demonstrated a complete end-to-end data science pipeline applied to a **real-world aerospace challenge**. From data collection and cleaning to visual analytics and predictive modelling, the study provided both **technical rigor** and **practical insights**.

Main conclusions:

* Payload mass, orbit type, and launch site are the most important predictors of landing success.
* Logistic Regression offered transparency, while SVM produced the best accuracy.
* Interactive dashboards and maps enhanced stakeholder engagement.

**Future Work**:

* Integrating real-time telemetry data for live predictions.
* Using deep learning for feature extraction from time-series flight data.
* Expanding predictions beyond landing success to **mission costs and payload outcomes**.

This project serves as both a case study in applied data science and a demonstration of how analytics can support the **reusable spaceflight revolution**.