Winning the Space Race with Data Science

SpaceX Falcon 9 Landing Prediction – End-to-End Data Science Project

Author: Chetan Sai Abhishek Injavarapu

Institution: IBM Applied Data Science Capstone (Coursera)

GitHub Repository: https://github.com/chetan-957/IBM-Capstone

1. Executive Summary

This project implements a complete end-to-end data science workflow to predict the successful landing of SpaceX Falcon 9 first-stage boosters. Reusable rocket technology has significantly reduced launch costs, but landing outcomes depend on multiple operational and technical factors.

The workflow included data collection via the SpaceX REST API, web scraping of landing outcomes, data wrangling, SQL-based exploratory data analysis, interactive visual analytics using Folium and Plotly Dash, and predictive modeling using classification algorithms.

All four classification models achieved approximately 83% test accuracy. The Decision Tree model demonstrated the highest cross-validation accuracy (~88.9%) and achieved perfect recall for successful landings.

2. Introduction

SpaceX revolutionized aerospace engineering through reusable Falcon 9 rockets. Successful recovery of first-stage boosters reduces costs and improves mission efficiency.

This report investigates key determinants of landing success, including payload mass, orbit type, launch site, booster version, and operational maturity.

Research Questions:

- What factors most influence landing success?
- Does payload mass impact landing probability?
- Do launch sites and orbit types affect outcomes?
- Has landing reliability improved over time?
- Can machine learning models accurately predict landing success?

3. Methodology

3.1 Data Collection

Launch data was collected using the SpaceX REST API, extracting flight number, payload mass, orbit type, launch site, booster version, and landing outcome. Landing history data was scraped from Wikipedia using BeautifulSoup and merged with API results into a structured dataset.

3.2 Data Wrangling

Data preprocessing included filtering Falcon 9 missions, removing irrelevant columns, handling missing values, encoding categorical variables, and creating a binary target variable representing landing success (1) or failure (0).

3.3 Exploratory Data Analysis

EDA was conducted using SQL and Python visualizations. Bar charts, scatter plots, and line charts were used to analyze launch site performance, payload trends, orbit complexity, and yearly success rates.

3.4 Interactive Analytics

Interactive maps were built using Folium to visualize geographic distribution and launch site proximity analysis. A Plotly Dash dashboard enabled dynamic filtering by launch site and payload range.

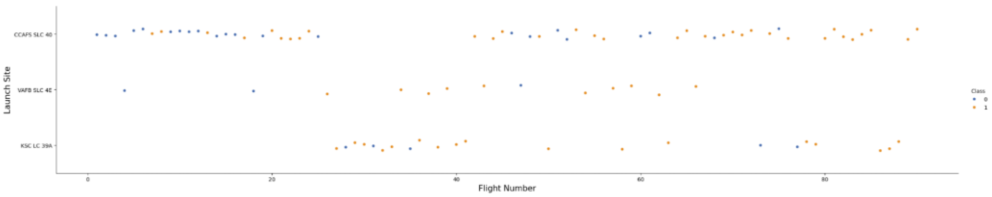
3.5 Predictive Modeling

Four classification models were implemented: Logistic Regression, Support Vector Machine, Decision Tree, and K-Nearest Neighbors. Data was split 80/20 for training and testing, and 10-fold cross-validation was used for hyperparameter tuning.

Analytical Results & Visual Insights

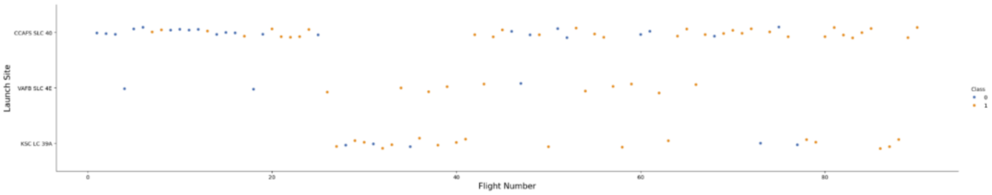
The following figures summarize key analytical insights derived from EDA, SQL analysis, geospatial modeling, interactive dashboards, and predictive classification.

Flight Number vs. Launch Site



•Strong Learning Curve Effect:
The increase in successful landings over flight number indicates operational improvement and engineering refinement.
•Experience Improves Reliability:
Later missions have significantly fewer failures, suggesting process stabilization and better booster reuse strategy.
•Launch Site Impact:
CCAFS handled early experimental phases, which explains more early failures.
KSC LC-39A appears during later, more mature stages, showing consistently strong success.
•Performance Evolution:
This pattern confirms that landing success is time-dependent and supports including Flight Number as a predictive feature.

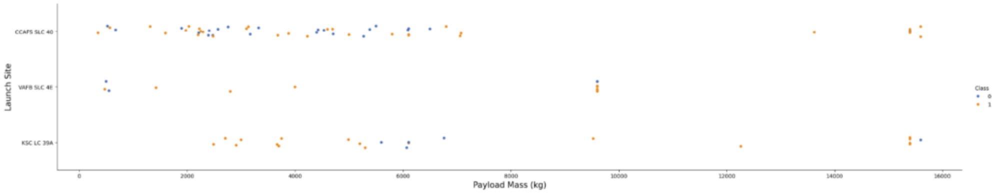
Flight Number vs. Launch Site



Conclusion

- Landing success improved significantly over time across all launch sites.
  The pattern clearly demonstrates technological advancement, operational learning, and increased reliability in later missions.

Payload vs. Launch Site



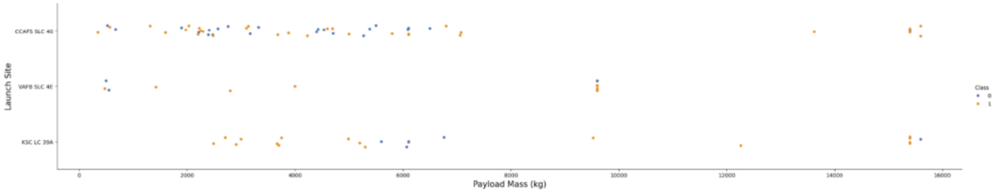
•Launch Site Specialization:
VAFB appears focused on lighter missions, possibly polar or specific orbital trajectories, which explains absence of heavy payload launches.

•Heavy Payload Capability:
KSC LC-39A and CCAFS SLC-40 handle heavier payloads, indicating stronger infrastructure and higher thrust mission profiles.

•Payload vs Success Pattern:
There is no clear linear decline in success with increasing payload mass.
In fact, later heavy payload missions show high success rates, suggesting improved booster performance.

•Operational Maturity Effect:
High-mass missions mostly appear in later flights, when landing technology had already matured — explaining strong success rates even for heavy payloads.

Payload vs. Launch Site

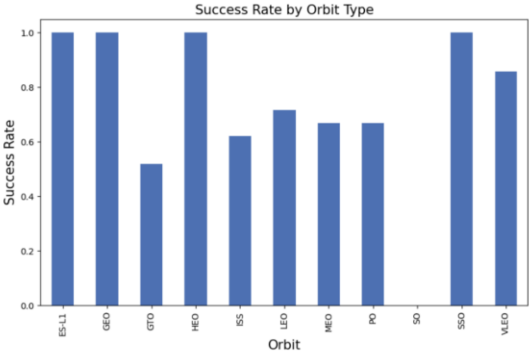


Conclusion

- Payload mass alone does not determine landing failure.
  Instead, landing success appears more strongly influenced by mission maturity and technological evolution rather than payload weight alone.

Success Rate vs. Orbit Type

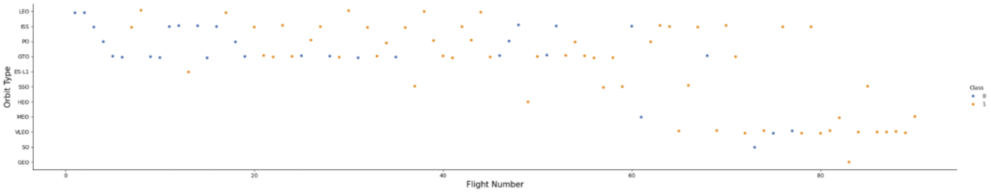
•Orbit Complexity Matters:
GTO missions typically require higher energy and more complex flight profiles. The lower success rate reflects higher landing difficulty.
•Mature Mission Profiles:
GEO and SSO show very high success rates, suggesting operational stability and well-optimized landing procedures for those trajectories.
•Operational Learning Curve:
More common orbits (LEO, ISS) show solid but not perfect success — indicating improvement over time but with earlier failures included in the data.
•Sample Size Warning:
Orbits like ES-L1 and HEO may show 100% success, but this likely reflects very few missions. We cannot over-generalize from small counts.



Conclusion:

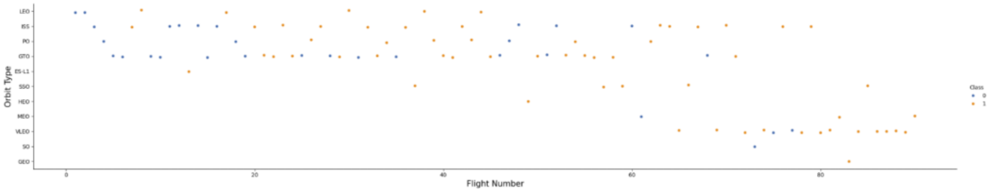
- Orbit type clearly influences landing success probability.
  Higher-energy orbits (like GTO) introduce greater landing risk, making Orbit a strong predictive feature for classification modeling.

Flight Number vs. Orbit Type



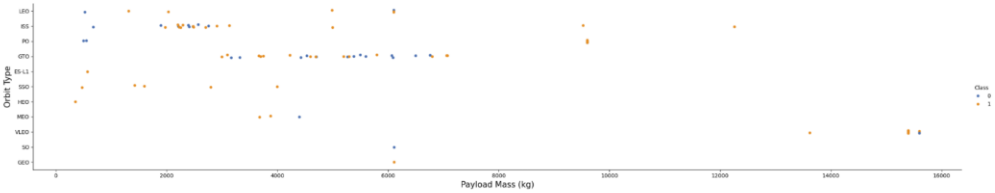
•Strong Learning Effect in LEO:
As flight number increases, success rate improves — clear evidence of operational maturity and landing optimization over time.
•Early Program Volatility:
Lower flight numbers show more failures across orbit types, consistent with early-stage testing and experimentation.
•Orbit-Specific Difficulty:
GTO does not show a strong upward trend in success over time, suggesting intrinsic mission complexity rather than just learning effects.
•Expansion into New Orbits:
Later flight numbers include newer orbit categories (VLEO, MEO) with higher success rates, indicating technological advancement and confidence.

Flight Number vs. Orbit Type



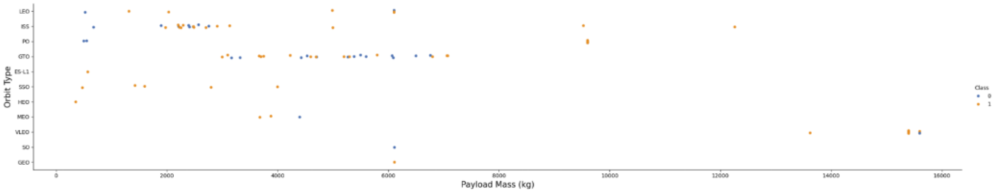
Conclusion:
•Flight number acts as a proxy for experience.
Landing success improves over time, particularly for common orbits like LEO.
This confirms that historical sequence (experience) is an important predictive feature for modeling success.

Payload vs. Orbit Type



•**Payload Alone Is Not Enough:**
Success does not depend purely on payload mass — orbit type clearly influences landing outcome.
•**LEO & ISS Are Operationally Stable:**
Even as payload mass increases, success remains consistently high — indicating optimized recovery procedures.
•**GTO Is More Challenging:**
At similar payload ranges, GTO shows both success and failure — suggesting mission complexity impacts landing reliability.
•**Heavy Payload Success in Later Missions:**
Very high payload missions appear later in the program and mostly succeed — again showing learning + technological advancement.

Payload vs. Orbit Type



Conclusion:

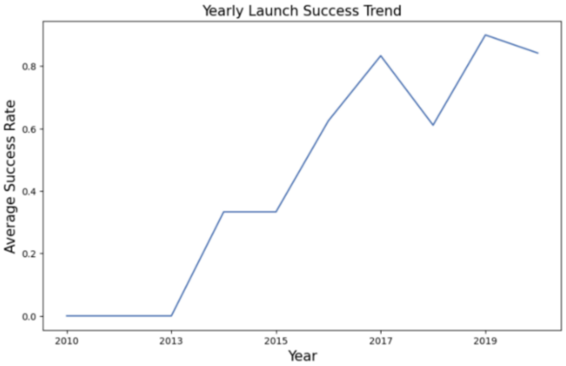
There is an interaction effect between **Payload Mass and Orbit Type**.
This justifies including both variables in the predictive model instead of treating them independently.

Launch Success Yearly Trend

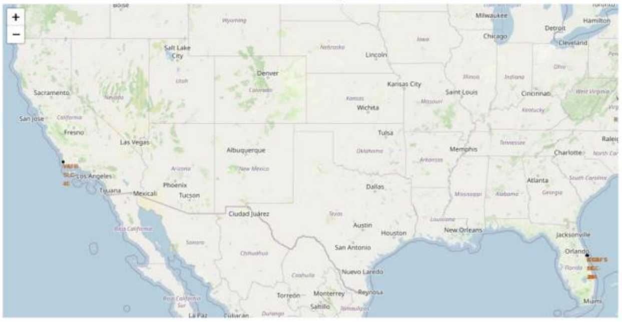
- **Clear Learning Curve Effect:**
  SpaceX significantly improved landing reliability over time.
- **Operational Maturity After 2015:**
  Post-2015 missions show consistent upward trend, indicating improved engineering and recovery systems.
- **Temporary Dip in 2018:**
  Slight drop suggests experimentation phase or mission complexity, but recovery was quick.
- **High Stability in Recent Years:**
  2019–2020 demonstrates mature reusable rocket technology.

Conclusion:

Time (Year / Flight Number) is a **strong predictive feature**.
Modeling should account for program maturity since later missions have higher probability of success.



Global Distribution of SpaceX Launch Sites



Important Elements
The map displays **all SpaceX launch site markers** plotted using Folium on a global geographic layout.

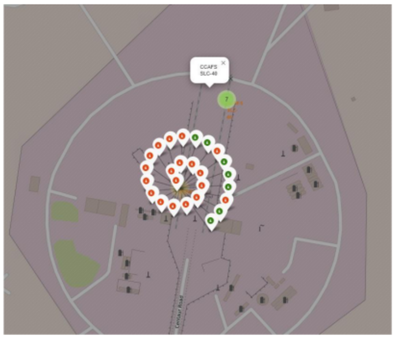
•Three primary locations are visible:

- **CCAFS LC-40** (Florida)
- **KSC LC-39A** (Florida)
- **VAFB SLC-4E** (California)

•The markers clearly show operations concentrated in **two U.S. coastal regions**: East Coast (Florida) and West Coast (California).

Key Findings
•**Bi-Coastal Strategy:** SpaceX operates from both Atlantic and Pacific coasts, allowing flexibility in orbital trajectories (LEO, ISS, polar orbits).
•**Florida Dominance:** Two launch pads are located in Florida, indicating it as the primary launch hub.
•**Strategic Geography:** Coastal locations reduce risk by allowing rocket stages to travel over ocean rather than populated areas.

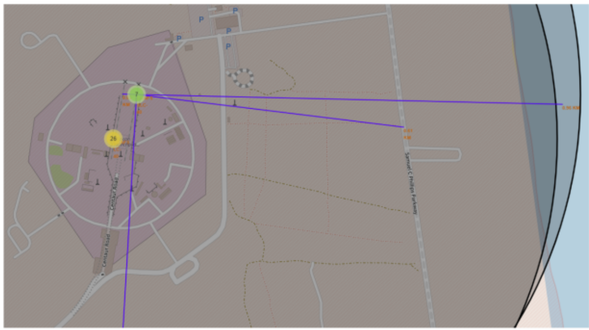
Color-Labeled Launch Outcomes at CCAFS LC-40



Important Elements
Each marker represents a single launch attempt at **CCAFS LC-40**.
•**Green markers = Successful Launch**
•**Red markers = Failed Launch**
•Marker clustering is enabled (numbered circle), showing high launch density at this site.

Key Findings
•Early launches show a higher concentration of **red markers**, indicating initial launching failures during the experimental phase.
•As missions progress, **green markers increase**, showing clear improvement in launching reliability.
•The dense clustering confirms that **CCAFS LC-40 is a primary operational hub**, playing a major role in SpaceX's learning curve and recovery optimization.

Proximity Analysis of Launch Site – CCAFS LC-40



Important Elements

The **launch site marker** is displayed with connecting blue distance lines to nearby infrastructure.

•Distances are clearly labeled on the map:

- **Railway ≈ 0.3 km**
- **Highway ≈ 0.6 km**
- **Coastline ≈ 0.9–1.0 km**

•The surrounding urban reference (Melbourne) is located much farther away (≈ **53.4 km**).

Key Findings

•**Close to Railway:**

The site's proximity (~0.3 km) supports efficient transportation of rocket components and heavy equipment.

•**Near Major Highway:**

Being ~0.6 km from a highway enables smooth logistics, workforce mobility, and emergency access.

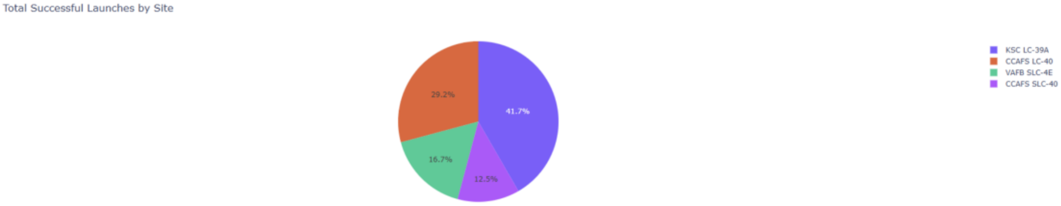
•**Coastal Location:**

At under 1 km from the coastline, launches can safely occur over the ocean, minimizing risk to populated land areas.

•**Far from Major Cities:**

The ~53 km distance from Melbourne confirms deliberate placement away from dense population centers for safety.

Launch Success Distribution Across All Sites



Key Highlights

•**KSC LC-39A leads** with the highest share of successful launches (~41.7%).

•**CCAFS LC-40** is second (~29.2%), showing strong performance.

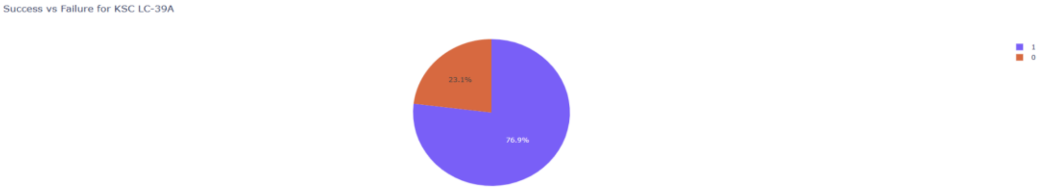
•**VAFB SLC-4E** and **CCAFS SLC-40** contribute smaller portions (~16.7% and ~12.5%).

•Success is concentrated mainly in two major launch sites.

Insight

KSC LC-39A is the dominant launch site in terms of total successful missions, indicating higher operational activity and strong performance compared to other sites.

KSC LC-39A – Highest Launch Success Ratio



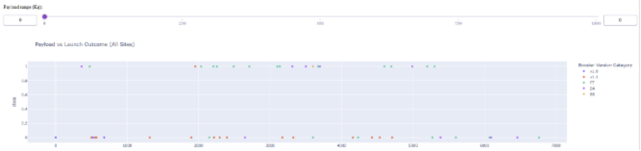
Key Highlights

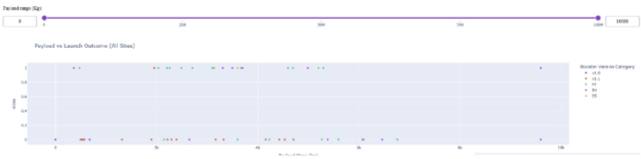
- **Success Rate: ~76.9%**
- **Failure Rate: ~23.1%**
- Majority of launches from KSC LC-39A are successful (Class = 1 dominates the chart).

Insight

KSC LC-39A not only has the highest number of successful launches but also the **highest success ratio**, making it the most reliable and best-performing launch site among all locations.

Payload Mass vs Launch Outcome Across All Sites





Overall Insights

- **Mid-range payloads (3000–6000 kg) have the highest success concentration.**
- **FT and B4 boosters demonstrate stronger performance compared to v1.0/v1.1.**
- Technological upgrades in booster versions correlate with improved success rates.

Screenshot 1: Narrow Payload Range (Low Range Selected)Key Observations:

- Lower payload ranges (0–3000 kg approx.) show **mixed outcomes**.
- Earlier booster versions (**v1.0 and v1.1**) show more failures (Class = 0).
- Success rate is less consistent in lower payload bands.

Screenshot 2: Full Payload Range (0–10000 kg Selected)Key Observations:

- Mid to higher payload ranges (~3000–6000 kg) show **more consistent success (Class = 1)**.
- **FT and B4 booster versions** dominate successful launches.
- Higher payloads (>7000 kg) appear limited but mostly successful.
- Earlier versions (v1.0, v1.1) show comparatively lower success rates.

Classification Accuracy

Objective

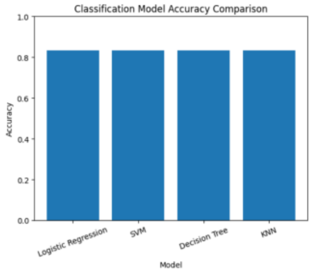
- To compare the performance of multiple classification models built to predict Falcon 9 first-stage landing success.

Models Evaluated

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K-Nearest Neighbors (KNN)

Results (Test Accuracy)

- All four tuned models achieved a similar test accuracy of approximately **83.33%**.



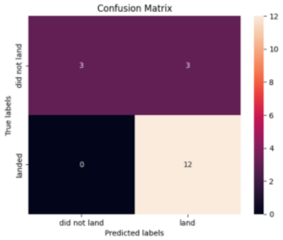
Best Performing Model

- Although test accuracy was identical across models, the **Decision Tree model achieved the highest cross-validation accuracy (~88.9%) during hyperparameter tuning**.
- Therefore, the **Decision Tree was selected as the best-performing model** based on overall training performance and validation stability.

Conclusion

- The models demonstrate comparable generalization performance; however, the **Decision Tree** shows stronger learning capability during cross-validation and is considered the most reliable model for this prediction task.

Confusion Matrix



Performance Insight

•The model is **very strong at predicting successful landings** (100% recall for landed cases).
•Slight weakness exists in distinguishing failed launches (some false positives).
•Overall test accuracy ≈ **83.33%**.

•The model correctly identified **all successful landings** (Recall = 1.00), meaning no successful missions were missed.
•Some failed launches were misclassified as successful (FP = 3), reducing precision to 0.80.
•The high F1 score (0.89) indicates strong overall balance between precision and recall.
•The model performs very well at detecting successful landings but is weaker at distinguishing failures.
This supports selecting the Decision Tree model as the best-performing classifier in this analysis.

Confusion Matrix Values
•True Positives (TP) = 12
•True Negatives (TN) = 3
•False Positives (FP) = 3
•False Negatives (FN) = 0
Total test samples = 18

The **Decision Tree** model performs reliably in predicting landing success, with strong success detection capability and moderate error in identifying failures. This makes it suitable for estimating mission landing outcomes while minimizing missed successful landings.

4. Results

Landing success improved significantly over time, demonstrating a strong learning curve effect.

KSC LC-39A emerged as the most reliable launch site, while GTO missions exhibited lower landing success due to higher mission complexity.

Model Performance:

- Test Accuracy (All Models): ~83.33%

- Best Model: Decision Tree

- Cross-Validation Accuracy: ~88.9%

- Confusion Matrix: TP=12, TN=3, FP=3, FN=0

- Recall (Success Detection): 1.00

- F1 Score: ~0.89

5. Discussion

The analysis confirms that booster version evolution, orbit complexity, payload interaction, and operational maturity significantly influence landing success.

Predictive Modeling Performance

All four models achieved comparable test accuracy (~83.33%). Decision Tree demonstrated the strongest cross-validation performance (~88.9%) and perfect recall (1.00) for successful landings.

Confusion Matrix Summary:
TP = 12
TN = 3
FP = 3
FN = 0

Model shows strong success detection capability with moderate false positives.

6. Conclusion

This project demonstrates how data science and machine learning can support aerospace decision-making and performance optimization. Predictive models can assist in mission risk assessment, cost estimation, and operational planning.

The findings validate that technological evolution and accumulated operational experience have significantly improved Falcon 9 landing reliability.

7. References

- SpaceX REST API Documentation
- Wikipedia Falcon 9 Launch Records
- Scikit-learn Documentation
- IBM Applied Data Science Capstone Labs