Winning the Space Race

Predicting Falcon 9 First Stage Landings

ROLE Lead Data Scientist, SpaceY

PROJECT Applied Data Science Capstone

Executive Summary

Data-Driven Proof of Reusability Feasibility

Main Objective

Determine the feasibility of reusing the Falcon 9 first stage to reduce SpaceY's launch costs and compete effectively with SpaceX.

Key Findings

Analysis confirms reusability is technically viable and predictable. Landing success is highly correlated with specific orbits and launch sites.

90%
Recent Success Rate

83.3%
Model Accuracy

Business Impact

The predictive model enables accurate cost estimation and identification of low-risk mission profiles, potentially saving millions per launch.

Introduction & Context

The Competitive Challenge of the Space Industry

SpaceX Cost	Competitors
$62M	**$165M+**
per launch	per launch

Business Problem

Savings come from reusability. To compete, SpaceY must predict if the first stage will land successfully based on pre-launch data.

01	02	03	04
Collect Data	Clean & Analyze	Model	Validate

Methodology

Data Collection & Wrangling Process

Data Sources

SpaceX REST API

Automated extraction of technical launch details: flight number, date, booster version, payload mass, orbit, and launch site.

Web Scraping

Historical data collection from Wikipedia ("List of Falcon 9 launches") using BeautifulSoup to obtain past mission outcomes.

Data Wrangling

Filtering & Cleaning

Exclusive focus on **Falcon 9** launches. Missing value treatment in PayloadMass by replacing with the mean (~6,123 kg).

Target Variable (Class)

Creation of a binary variable for classification:

1 = Successful Landing

0 = Unsuccessful Landing (Crash, Ocean, None)

Methodology

EDA & Interactive Visual Analytics

Exploratory Data Analysis (EDA)

Extracting critical metrics and visualizing relationships between variables such as Flight Number, Orbit Type, and Payload Mass.

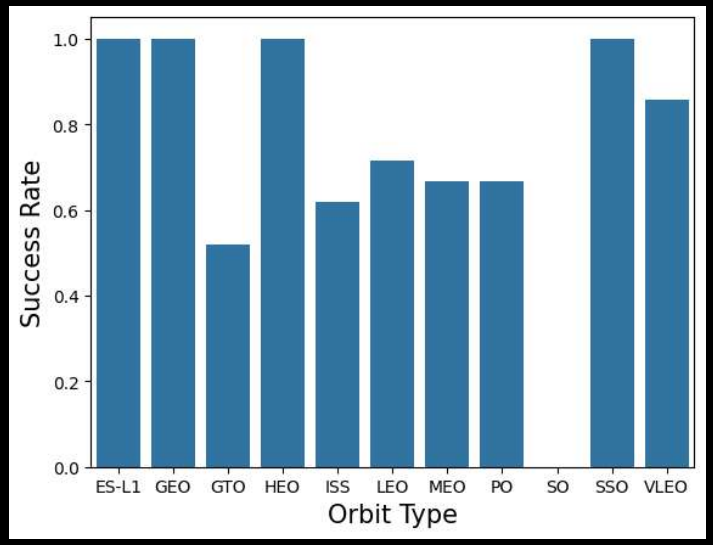
SQL | Pandas | Seaborn

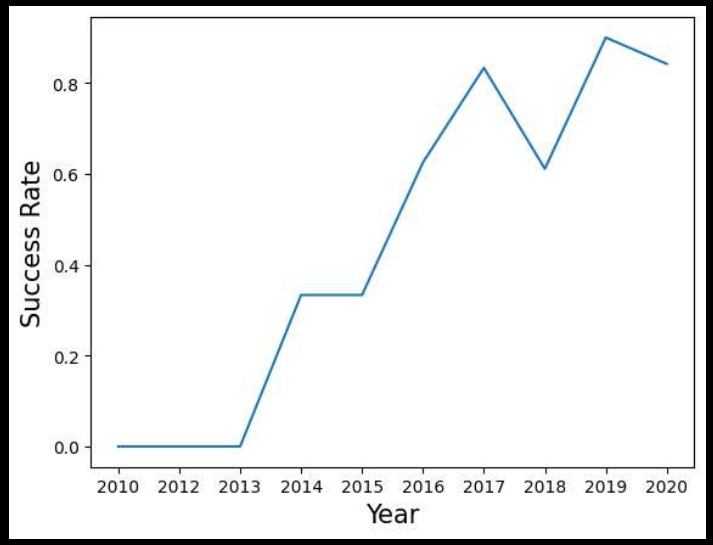
Interactive Analytics

Geospatial mapping of launch sites and dynamic dashboards for real-time success rate filtering.

Folium Maps | Plotly Dash

> *"Identify visual patterns and hidden correlations that inform the construction of robust predictive models."*





Methodology: Predictive Analysis

Machine Learning Models and Optimization

Data Preprocessing

- **Standardization:** Used StandardScaler to normalize features.
- **Data Splitting:** 80% Training | 20% Testing for robust validation.

Models Tested

Logistic Regression
Baseline for binary classification.

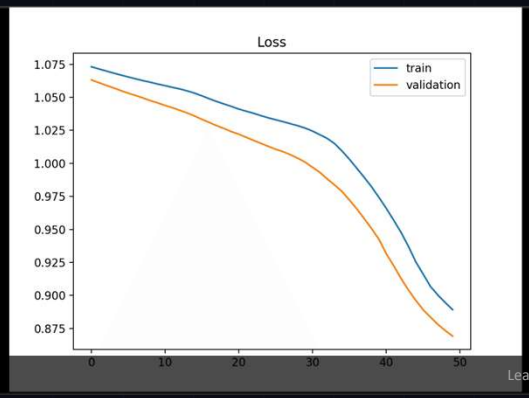
SVM
RBF kernel for non-linear boundaries.

Decision Tree
Interpretable rule-based model.

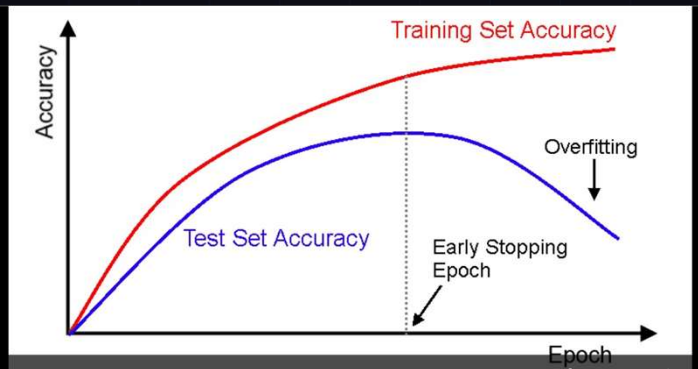
KNN
Proximity-based classification.

Optimization & Evaluation

- **GridSearchCV:** Exhaustive hyperparameter tuning.
- **Cross-Validation:** 10-fold CV implementation.
- **Metrics:** Accuracy, Confusion Matrix, Precision, and Recall.



Learning Curves Analysis



Performance Metrics Deep Dive

EDA Results

Visual Insights & Patterns

Payload vs. Launch Site

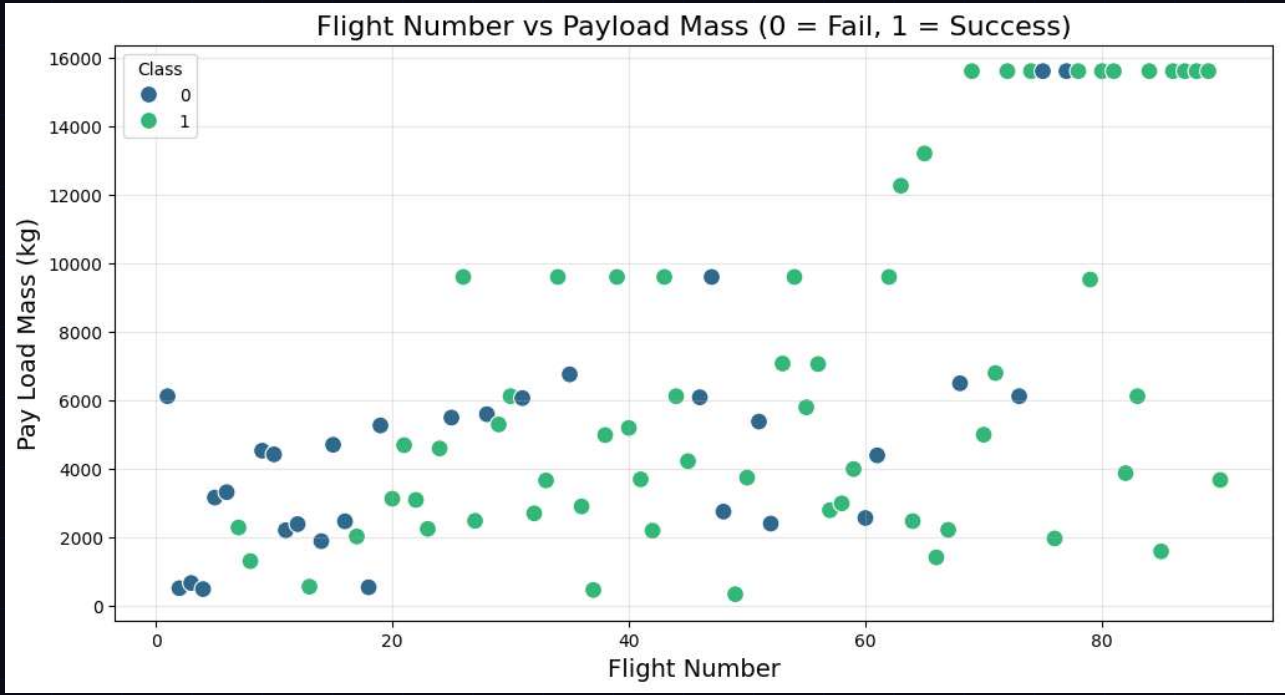
Heavy payloads (>8,000kg) are exclusively launched from **KSC LC-39A** and **CCSFS SLC 40**. VAFB SLC 4E handles lighter polar orbit missions.

Success Rate by Orbit

100% Success Rate observed for ES-L1, GEO, HEO, and SSO orbits. SO (Sun-Synchronous) orbits showed the lowest performance.

Yearly Trend

Consistent growth in success rates from 2013 to 2020, validating the technological maturity of the landing system over time.



EDA Results: SQL Insights

Key Statistics Extracted via SQL

45,596 kg

Total NASA Payload

CRS Missions

2,534 kg

Average Payload

F9 v1.1

98

Successful Missions

Recorded in Database

SLC 40

Top Launch Site

CCSFS (Most Frequent)

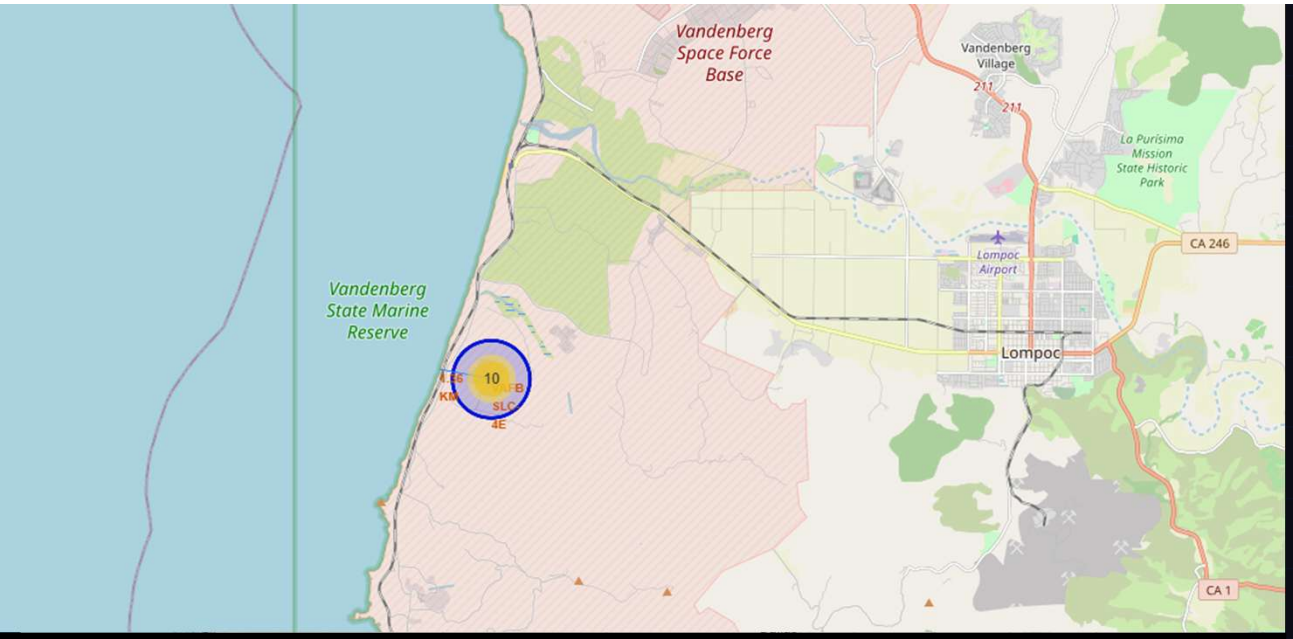
15,600 kg

Max Payload Booster

F9 B5 Series (e.g., B1048)

Strategic Insight

High-activity sites like CCSFS SLC 40 and KSC LC-39A demonstrate robust infrastructure and a proven track record, making them priority candidates for SpaceY development.



Results

Geospatial Analysis & Safety

Coastal Proximity

All launch sites are strategically located near the coastline to facilitate oceanic recovery.

VAFB SLC-4E: ~1.36 KM from coast

Flight Corridors

Launches are directed over the ocean to minimize risk to populated areas, cities, and railways.

Success Clusters

Cluster visualization confirms that successful landings vastly outnumber failures in designated zones.

Recommendation

"SpaceY should prioritize the development of coastal facilities to maximize safety and align with proven industry success standards."

Results: Plotly Dashboard

Dynamic Insights on Payload and Performance

Dashboard Capabilities

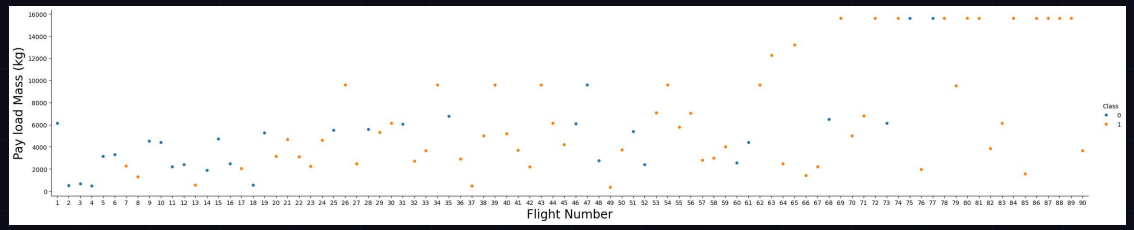
- Dynamic filtering by **Launch Site**
- Input slider to filter by **Payload Mass (kg)**
- Interactive real-time visualization of success rates

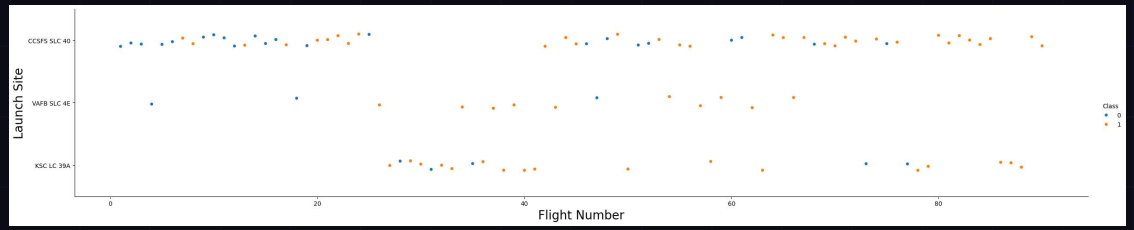
Main Insight

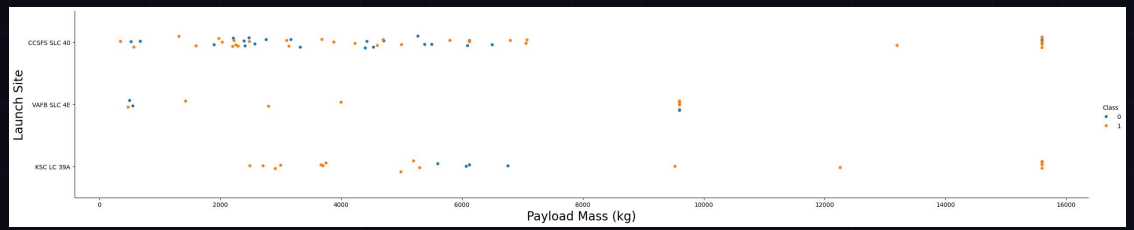
Mid-range payloads (2,000kg - 6,000kg) show the highest consistency of landing success across all sites.

Practical Application

Quick mission feasibility assessment based on payload specifications and launch site.







Results

Machine Learning Model Performance

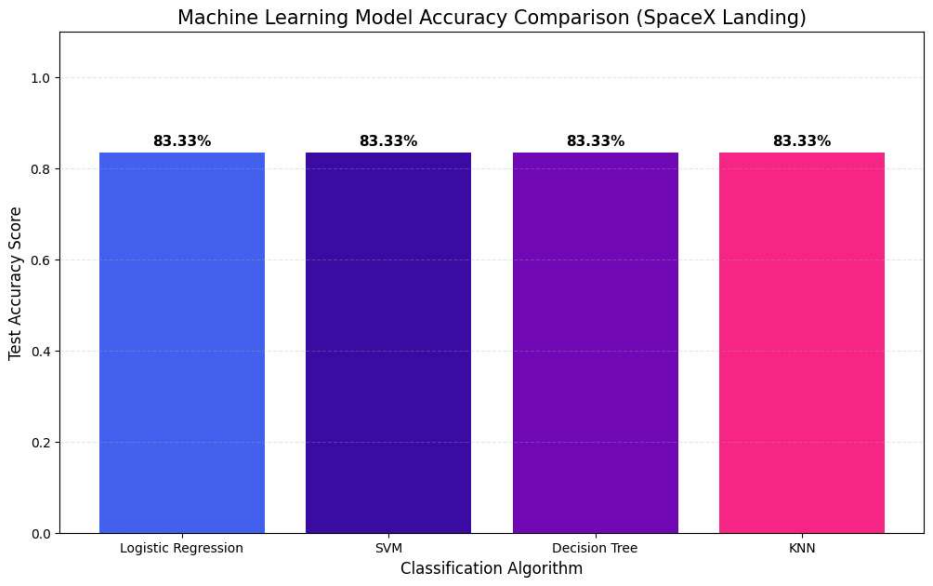
Logistic Regression	**83.33%**
Support Vector Machines (SVM)	**83.33%**
K-Nearest Neighbors (KNN)	**83.33%**
Decision Tree	**72.22%**

Confusion Matrix Analysis

Models predict "Success" (1) with high precision. The primary source of error is False Positives influenced by class imbalance.

Recommendation

We recommend using **Logistic Regression** or **SVM** for their consistency and lower risk of overfitting.



Conclusion

Strategic Path for SpaceY

Key Conclusions

- Using classification models (DT, SVM, KNN). **~83% accuracy**
- **Orbit Type** and **Payload Mass** are the most critical determinants of success.
- Coastal launch infrastructure is the proven industry standard for safety and recovery.

Strategic Recommendations

1. Infrastructure

Prioritize coastal launch facilities similar to KSC LC-39A.

2. Market Focus

Initially concentrate on LEO and GTO missions (high success rates).

3. Risk Management

Use predictive models to assess risk before reuse contracts.

4. Optimization

Focus on mid-range payloads (2k-6k kg) for maximum consistency.

Next Steps: Refine model with real-time data and integrate into the operational decision pipeline.