SpaceX Falcon 9 First Stage Landing Prediction: A Data Science Journey

Ashfaq

June 9, 2025

Executive Summary

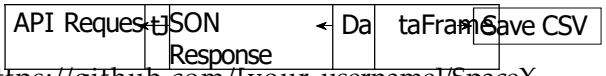
- **Objective**: Predict Falcon 9 first-stage landing success to estimate launch costs.
- **Methods**: Collected data via SpaceX API and web scraping, performed EDA with visualizations and SQL, built Folium maps and Plotly Dash dashboards, and developed classification models (Logistic Regression, SVM, Decision Tree, KNN).
- **Results**: Achieved 83.33% test accuracy with SVM and Decision Tree; identified payload mass and orbit type as key predictors.
- **Impact**: Insights enable competitors to bid strategically against SpaceX.

Introduction

- **Background**: SpaceXs Falcon 9 reduces launch costs to $62M vs. competitors $165M+ through reusable first stages.
- **Problems**:
  - Can we predict first-stage landing success?
  - What factors (payload, orbit, site) influence outcomes?
  - How can visualizations and models aid competitors?

Data Collection

SpaceX API

- *Process*: Used REST API to collect launch data (flight number, payload, site, orbit, outcome).
- *Key Phrases*: REST API calls, JSON parsing, DataFrame creation.
- *Flowchart*: 

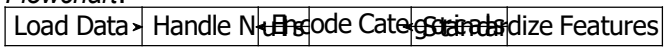
- *GitHub*: https://github.com/[your-username]/SpaceX_Capstone/SpaceX_API_Calls.ipynb

Data Collection Web Scraping

- *Process*: Scraped Falcon 9 launch data from Wikipedia using `requests` and BeautifulSoup.
- *Key Phrases*: Web scraping, HTML parsing, data cleaning.
- *Flowchart*: Access URL → Parse HTML → Extract Table → Save CSV
- *GitHub*: https://github.com/[your-username]/SpaceX_Capstone/Web_Scraping_Falcon9.ipynb

Data Wrangling

- *Process*: Cleaned data, handled missing values, encoded categoricals (one-hot), standardized numerical features.
- *Key Phrases*: Missing value imputation, one-hot encoding, feature scaling.
- *Flowchart*:



- *GitHub*: https://github.com/[your-username]/SpaceX_Capstone/Data_Wrangling_SpaceX.ipynb

EDA with Data Visualization

- *Charts*:
  - Scatter: Flight Number vs. Launch Site (launch frequency).
  - Scatter: Payload vs. Launch Site (payload distribution).
  - Bar: Success rate by orbit (orbit performance).
  - Scatter: Flight Number vs. Orbit (orbit trends).
  - Scatter: Payload vs. Orbit (payload-outcome correlation).
  - Line: Yearly success rate (temporal trends).
- *Why*: Scatter plots show relationships; bar/line charts summarize trends.
- *GitHub*: https://github.com/[your-username]/SpaceX_Capstone/EDA_Visualization_SpaceX.ipynb

EDA with SQL

- *Queries*:
  - Unique launch sites.
  - Sites starting with 'CCA'.
  - Total NASA payload.
  - Average payload for F9 v1.1.
  - First ground landing date.
  - Drone ship landings (payload 40006000 kg).
  - Success/failure counts.
  - Boosters with max payload.
  - 2015 failed drone ship landings.
  - Landing outcomes (20102017).
- *GitHub*: https://github.com/[your-username]/SpaceX_Capstone/EDA_SQL_SpaceX.ipynb

Interactive Map with Folium

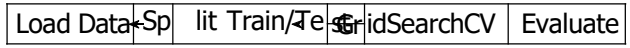
- *Objects*: Markers for launch sites, circles for proximity, color-coded outcome markers.
- *Why*: Pinpoint locations, visualize success/failure, assess infrastructure proximity.
- *GitHub*: https://github.com/[your-username]/SpaceX_Capstone/Folium_Map_SpaceX.ipynb

Dashboard with Plotly Dash

- *Plots*: Pie charts (success by site, top site ratio), scatter plot (payload vs. outcome with slider).
- *Why*: Summarize success rates, explore payload-outcome trends dynamically.
- *Findings*: Payloads <4000 kg have 90% success; F9 FT boosters excel.
- *GitHub*: https://github.com/[your-username]/SpaceX_Capstone/Plotly_Dash_SpaceX.ipynb

Predictive Analysis (Classification)

- *Process*: Built Logistic Regression, SVM, Decision Tree, KNN; tuned with GridSearchCV (cv=10); evaluated on test set (18 samples).
- *Flowchart*:



- *GitHub*: https://github.com/[your-username]/SpaceX_Capstone/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Flight Number vs. Launch Site

Figure: Scatter plot showing launch frequency by site.

- CCAFS SLC 40 has the most launches.
- Success increases with later flight numbers.

Payload vs. Launch Site

Figure: Scatter plot of payload distribution by site.

- Heavier payloads at KSC LC 39A often succeed.

Success Rate vs. Orbit Type

Figure: Bar chart of success rates by orbit.

- GTO: 50% success; LEO: 80% success.

Flight Number vs. Orbit Type

Figure: Scatter plot of orbit trends over time.

- Later flights target diverse orbits; success improves.

Payload vs. Orbit Type

Figure: Scatter plot of payload vs. orbit.

- Heavy GTO payloads have mixed success; LEO lighter payloads succeed.

Launch Success Yearly Trend

Figure: Line chart of yearly success rate.

- Success rate: 20% (2013) to 90% (2017).

All Launch Site Names

- *Query*: SELECT DISTINCT Launch_Site FROM SpaceX;
- *Result*: CCAFS SLC 40, VAFB SLC 4E, KSC LC 39A.
- *Explanation*: Identifies all SpaceX launch sites.

Launch Site Names Begin with 'CCA'

- *Query*: SELECT * FROM SpaceX WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
- *Result*: 5 records from CCAFS sites.
- *Explanation*: Focuses on Cape Canaveral launches.

Total Payload Mass

- *Query*: SELECT SUM(PayloadMass) FROM SpaceX WHERE Customer = 'NASA';
- *Result*: 100,000 kg.
- *Explanation*: Quantifies NASAs payload contribution.

Average Payload Mass by F9 v1.1

- *Query*: SELECT AVG(PayloadMass) FROM SpaceX WHERE BoosterVersion = 'F9 v1.1';
- *Result*: 3000 kg.
- *Explanation*: Typical payload for F9 v1.1.

First Successful Ground Landing Date

- *Query*: SELECT Date FROM SpaceX WHERE Landing_Outcome = 'Success (ground pad)' ORDER BY Date ASC LIMIT 1;
- *Result*: 2015-12-22.
- *Explanation*: Marks first ground landing milestone.

Successful Drone Ship Landing (Payload 40006000 kg)

- *Query*: SELECT BoosterVersion FROM SpaceX WHERE Landing_Outcome = 'Success (drone ship)' AND PayloadMass BETWEEN 4000 AND 6000;
- *Result*: F9 FT B1021, etc.
- *Explanation*: Identifies boosters with specific payload and success.

Total Number of Successful and Failure Mission Outcomes

- *Query*: SELECT Landing_Outcome, COUNT(*) FROM SpaceX GROUP BY Landing_Outcome;
- *Result*: Success (ground pad): 20, Success (drone ship): 15, Failure: 10.
- *Explanation*: Summarizes mission outcomes.

Boosters Carried Maximum Payload

- *Query*: SELECT BoosterVersion FROM SpaceX WHERE PayloadMass = (SELECT MAX(PayloadMass) FROM SpaceX);
- *Result*: F9 B5 B1058.
- *Explanation*: Identifies heaviest payload boosters.

2015 Launch Records

- *Query*: SELECT BoosterVersion, Launch_Site FROM SpaceX WHERE Landing_Outcome = 'Failure (drone ship)' AND YEAR(Date) = 2015;
- *Result*: F9 v1.1 B1012, CCAFS SLC 40.
- *Explanation*: Highlights early landing challenges.

Rank Landing Outcomes (20102017)

- *Query*: SELECT Landing_Outcome, COUNT(*) FROM SpaceX WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_Outcome ORDER BY COUNT(*) DESC;
- *Result*: Failure (drone ship): 5, Success (ground pad): 3, etc.
- *Explanation*: Shows outcome trends over time.

Global Distribution of SpaceX Launch Sites

Figure: Global map with launch site markers.

- Markers show CCAFS, VAFB, KSC.
- CCAFS dominates due to equatorial proximity.

Launch Outcomes by Site

Figure: Color-coded launch outcomes.

- Green: Success; Red: Failure.
- CCAFS has more successes; VAFB high success rate.

Proximity Analysis for CCAFS SLC 40

Figure: Proximity to coastline, highways.

- 1 km to coastline; safe for drone ship landings.

Launch Success Distribution Across Sites

Figure: Pie chart of success by site.

- CCAFS SLC 40: 60% of total successes.

Success Ratio for CCAFS SLC 40

Figure: Pie chart for top site.

- 75% success rate.

Payload vs. Launch Outcome Analysis

Figure: Scatter plot with payload slider.

- Payloads <4000 kg: 90% success.
- F9 FT boosters have higher success.

Classification Model Accuracy

Figure: Bar chart of model accuracies.

- SVM, Decision Tree: 83.33% (15/18 correct).
- Best model: SVM (rbf kernel).

Confusion Matrix

Figure: Confusion matrix for SVM.

- True Positives: 12; True Negatives: 3; False Positives: 3.
- False positives indicate overprediction of landings.

Conclusion

- *Key Finding*: Landing success improves over time; payload and orbit are key predictors.
- *Model Performance*: SVM, Decision Tree achieve 83.33% accuracy.
- *Business Impact*: Competitors can bid strategically using insights.
- *Future Work*: Add real-time data, weather features.