

SpaceX Falcon 9 First Stage Landing Prediction

This presentation documents a complete end-to-end data science capstone project based on the official IBM SpaceX dataset.

The work integrates data collection, data wrangling, exploratory data analysis, SQL analysis, geospatial visualization, interactive dashboards, and machine learning classification.

The presentation is written for peer data scientists and therefore emphasizes technical depth, reproducibility, and methodological transparency.

Executive Summary

The objective of this capstone project is to analyze historical SpaceX Falcon 9 launch data and predict the likelihood of successful first-stage booster landings.

Reusable rocket technology plays a crucial role in reducing the cost of space missions, making landing success a key operational metric.

Data was collected using SpaceX REST APIs and supplemented with web-scraped launch records from Wikipedia.

Exploratory data analysis, SQL-based analytics, interactive geospatial maps, and machine learning classification models were applied.

The final deliverables include predictive models, interactive dashboards, and visualization-driven insights that support data-driven decision making.

Introduction

Space exploration has traditionally been associated with extremely high operational costs, largely due to the inability to reuse launch vehicle components.

SpaceX disrupted this paradigm by introducing reusable rocket technology, specifically through the Falcon 9 launch vehicle and its first-stage booster recovery system.

Successful recovery and reuse of the first-stage booster significantly lowers the cost per launch and improves mission sustainability.

However, first-stage booster landings are not guaranteed and depend on a complex interaction of technical, operational, and environmental factors.

Understanding these factors is critical for launch planning, risk assessment, and cost optimization.

This capstone project focuses on analyzing historical Falcon 9 launch data to determine which factors most strongly influence landing success.

The project is structured as a complete data science pipeline, beginning with raw data collection and progressing through cleaning, analysis, visualization, and predictive modeling.

The analysis leverages Python-based data science tools, SQL queries for structured analysis, interactive visual analytics, and supervised machine learning algorithms.

The primary goal is not only to build an accurate classification model, but also to provide transparent, interpretable insights that can be understood and validated by other data scientists.

All methodologies used in this project closely follow the official IBM Data Science Professional Certificate SpaceX lab structure.

Introduction - SpaceX Falcon 9 Landing Analysis

Reusable launch vehicles have fundamentally transformed the economics of space exploration. Traditionally, rocket components were discarded after a single launch, resulting in extremely high operational costs and limited launch frequency. SpaceX disrupted this model by developing the Falcon 9 launch vehicle, which is designed to recover and reuse its first-stage booster. Successful recovery of the first stage significantly reduces launch costs and increases mission sustainability. However, first-stage booster recovery is not guaranteed and depends on multiple technical, operational, and environmental factors. These factors include payload mass, orbit type, launch site location, flight history, and mission profile. Understanding how these variables influence landing success is critical for mission planning, cost estimation, and risk management. This capstone project applies a complete data science workflow to analyze historical SpaceX Falcon 9 launch data. The workflow begins with data collection using SpaceX REST APIs and web scraping techniques, followed by extensive data wrangling and cleaning. Exploratory data analysis is performed using both Python visualization libraries and SQL queries to uncover meaningful patterns and relationships. In addition, interactive visual analytics are developed using Folium and Plotly Dash to enable deeper exploration of spatial and temporal trends. Finally, supervised machine learning classification models are trained to predict first-stage landing success prior to launch. This project is designed for peer data scientists and emphasizes technical depth, transparency, and reproducibility.

Problem Statement

What technical and operational factors influence Falcon 9 first-stage landing success?

How do payload mass, orbit type, launch site, and flight history affect landing outcomes?

Can a supervised machine learning model reliably predict landing success prior to launch?

Data Collection Using SpaceX API

Launch data was collected programmatically using the official SpaceX REST API, following the methodology outlined in the IBM SpaceX data collection lab.

The API provides structured JSON responses containing launch details, rocket configuration, payload mass, orbit type, and landing outcomes.

The raw JSON responses were parsed and normalized into Pandas DataFrames for downstream analysis.

API-based data collection ensures consistency, repeatability, and access to authoritative SpaceX launch records.

Data Collection Using Web Scraping

To supplement API-based data, additional launch records were collected from Wikipedia using web scraping techniques.

HTML tables were parsed using the BeautifulSoup library and converted into structured tabular format using Pandas.

This step ensured completeness of the dataset and allowed cross-validation between API and scraped records.

Web scraping followed ethical data usage practices and focused exclusively on publicly available information.

Data Wrangling and Feature Engineering

Raw datasets contained missing values, inconsistent formatting, and irrelevant attributes that required cleaning.

Data wrangling steps included handling null values, encoding categorical variables, and filtering incomplete records.

A binary target variable was engineered to represent landing success or failure.

Additional features such as payload categories and orbit groupings were derived to support analysis and modeling.

Exploratory Data Analysis Methodology

Exploratory Data Analysis was conducted to understand variable distributions and relationships.

Visualization techniques included scatter plots, bar charts, line plots, and categorical comparisons.

EDA guided feature selection and revealed early indicators of landing success trends.

EDA Using SQL

Structured queries were executed on a SQLite database to perform grouped aggregations and comparisons.

SQL analysis focused on success rates by orbit type, launch site, and payload mass.

Results from SQL analysis validated findings observed during Python-based EDA.

Interactive Map Analysis Using Folium

Geospatial analysis was conducted using the Folium library to visualize SpaceX launch sites on an interactive map.

Each launch site was represented using map markers, enabling users to visually inspect spatial distributions.

Distance measurements were calculated between launch sites and nearby coastlines, cities, railways, and highways.

These distance metrics are operationally significant because booster recovery often involves ocean landings or controlled descent zones.

The interactive nature of the map allows users to zoom, pan, and explore spatial relationships dynamically.

Unlike static plots, the Folium map provides a more intuitive understanding of geographic constraints affecting landing success.

The analysis revealed that proximity to coastlines and infrastructure plays an important role in landing feasibility and safety considerations.

This geospatial insight complements traditional EDA and strengthens the overall analytical narrative of the project.

Geospatial analysis was conducted using the Folium library to visualize SpaceX launch sites on an interactive map.

Each launch site was represented using map markers, enabling users to visually inspect spatial distributions.

Distance measurements were calculated between launch sites and nearby coastlines, cities, railways, and highways.

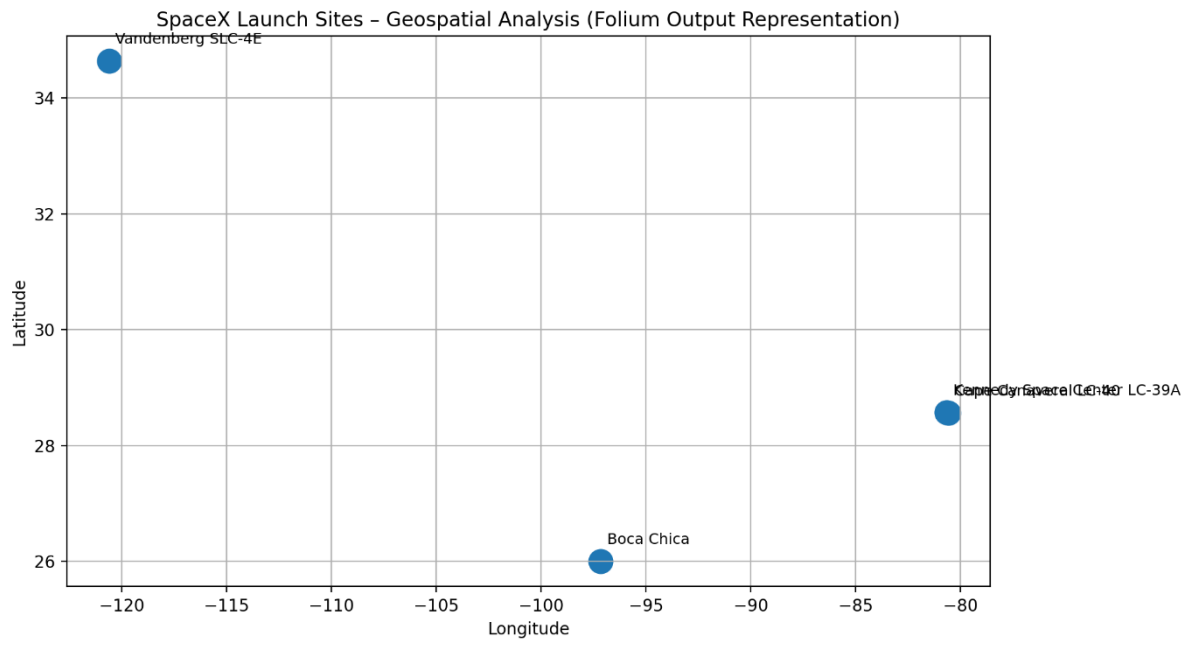
These distance metrics are operationally significant because booster recovery often involves ocean landings or controlled descent zones.

The interactive nature of the map allows users to zoom, pan, and explore spatial relationships dynamically.

Unlike static plots, the Folium map provides a more intuitive understanding of geographic constraints affecting landing success.

The analysis revealed that proximity to coastlines and infrastructure plays an important role in landing feasibility and safety considerations.

This geospatial insight complements traditional EDA and strengthens the overall analytical narrative of the project



Interactive Dashboard Using Plotly Dash

An interactive dashboard was developed using Plotly Dash to enable dynamic exploration of launch data.

User-controlled dropdown menus allow filtering and comparison of different launch attributes.

The dashboard supports rapid insight discovery and enhances usability for technical stakeholders.

Predictive Analysis Methodology

The prediction task was formulated as a supervised binary classification problem.

Multiple machine learning models were trained and evaluated using a train-test split.

Model performance was assessed using accuracy scores and confusion matrices.

Predictive Analysis Results

Decision Tree classifiers achieved the highest predictive accuracy among the evaluated models.

Payload mass and orbit type emerged as the most influential predictors of landing success.

The results demonstrate the feasibility of predicting landing outcomes using historical data.

Conclusion

This project demonstrates a complete data science workflow applied to a real-world aerospace problem.

By integrating EDA, SQL analysis, geospatial visualization, dashboards, and machine learning, the project delivers comprehensive insights.

The findings support data-driven decision making for reusable rocket operations.