SpaceX Falcon 9 First Stage Landing Prediction

This presentation documents the complete end-to-end data science capstone project focused on SpaceX Falcon 9 launches.

The work is intended for peer data scientists and therefore includes detailed technical explanations, methodologies, and results.

Executive Summary

This capstone project applies a full data science lifecycle to analyze SpaceX Falcon 9 launch data with a focus on first-stage booster landing outcomes.

Multiple data sources were combined using API-based data collection and web scraping techniques to ensure completeness and accuracy.

Exploratory data analysis, SQL-based analysis, interactive visual analytics, and predictive machine learning models were used.

The final outcome is a classification model capable of predicting landing success, supported by interactive dashboards and maps.

Introduction

Reusable launch vehicles represent a major advancement in the aerospace industry by significantly reducing the cost of space missions.

SpaceX’s Falcon 9 rocket achieves reusability by successfully landing its first-stage booster after launch, which can then be refurbished and reused.

However, not every mission results in a successful landing. Predicting whether a first-stage booster will land successfully is therefore a critical problem.

This project aims to analyze historical Falcon 9 launch data to identify the factors that most strongly influence landing success.

The analysis is designed for technical audiences and emphasizes transparency, reproducibility, and rigorous data-driven reasoning.

Problem Statement

Identify the launch and payload characteristics that influence Falcon 9 first-stage landing success.

Analyze how orbit type, payload mass, and launch site affect the probability of successful landings.

Develop a predictive classification model to estimate landing success before launch.

Data Collection – API

Launch data was collected from the official SpaceX REST API, which provides structured and reliable access to historical launch information.

The API responses were returned in JSON format and included details such as rocket configuration, payload mass, orbit type, and launch outcomes.

The raw JSON data was parsed and converted into Pandas DataFrames for further processing and analysis.

Data Collection – Web Scraping

Additional launch data was collected using web scraping techniques from publicly available Wikipedia tables.

BeautifulSoup was used to parse HTML content and extract relevant tables containing launch outcomes.

This scraped data was combined with API data to improve completeness and cross-validate launch records.

Data Wrangling and Cleaning

The collected datasets contained missing values and inconsistent formats that required cleaning.

Data wrangling steps included handling null values, standardizing categorical variables, and filtering irrelevant records.

A binary target variable was created to represent landing success or failure, enabling classification modeling.

EDA Methodology

Exploratory Data Analysis was conducted to understand data distributions and relationships between variables.

Visual techniques such as scatter plots, bar charts, and line plots were used to identify trends and anomalies.

EDA provided key insights that guided feature selection for machine learning models.

EDA with Visualization Results

The success rate of Falcon 9 landings has increased over time as SpaceX refined its technology.

Lower payload masses were observed to have higher probabilities of successful landings.

Certain orbit types such as LEO and ISS were associated with higher landing success rates.

EDA with SQL Results

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

Launch success rates were analyzed by orbit type, launch site, and payload category.

The SQL-based analysis confirmed trends identified during Python-based EDA.

Interactive Map with Folium

An interactive map was created using the Folium library to visualize SpaceX launch sites geographically.

Markers were added to represent launch locations, and distance measurements were calculated relative to coastlines, cities, and infrastructure.

The map enabled interactive exploration of spatial relationships that are difficult to observe in static charts.

Geospatial analysis revealed that proximity to coastlines and safety zones plays an important role in landing feasibility.

Plotly Dash Dashboard

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

Dropdown menus enable users to filter data and visualize different aspects of launch performance.

The dashboard enhances usability and supports data-driven decision-making.

Predictive Analysis Methodology

The prediction task was formulated as a binary classification problem.

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

Model performance was compared using accuracy and confusion matrices.

Predictive Analysis Results

Among the tested models, the Decision Tree classifier achieved the highest accuracy.

Payload mass and orbit type emerged as the most influential features.

The results demonstrate that landing success can be predicted with reasonable accuracy.

Innovative Insights

Combining geospatial analysis with machine learning provided deeper insights than traditional EDA alone.

The integration of SQL, visualization, and predictive modeling strengthened result reliability.

Conclusion

This project demonstrates the application of a complete data science workflow to a real-world aerospace problem.

The findings provide actionable insights into factors affecting Falcon 9 landing success.

The project highlights the value of data-driven approaches in complex engineering domains.