Project Report: SpaceX Falcon 9 Landing Prediction

Author: George Monappallil

Date: December 16, 2025

1. Executive Summary

The commercial space industry has undergone a paradigm shift driven by SpaceX's introduction of the Falcon 9 reusable rocket. This innovation has drastically reduced space access costs, lowering the price of a launch from over $165 million for expendable rockets to approximately $62 million for reusable ones.
The primary objective of this project was to predict whether the Falcon 9 first stage would land successfully based on historical flight data. By successfully predicting landing outcomes, competitors and stakeholders can determine the true cost of a launch for strategic bidding. The project utilized a multi-faceted approach involving data collection via APIs and web scraping, exploratory data analysis (EDA) using SQL and visualization tools, and predictive modeling using classification algorithms.
The analysis confirmed that rocket reusability has transitioned from an experimental phase to a predictable operation. Predictive models achieved an accuracy of 83.33%, demonstrating that landing outcomes can be reliably forecasted using flight features.

2. Methodology

2.1 Data Collection

To build a robust dataset, data was gathered from two primary sources:

- **SpaceX REST API:** Real-time launch data was extracted using `GET` requests to the endpoint `api.spacexdata.com/v4/launches/past`. The extraction parsed raw JSON responses to retrieve key features such as Payload Mass, Orbit, Booster Version, and Launch Site.
- **Web Scraping:** To supplement API gaps, historical launch records were scraped from the Wikipedia page "List of Falcon 9 and Falcon Heavy launches" using `BeautifulSoup`. This process extracted tabular data regarding launch dates, outcomes, and payload details.

2.2 Data Wrangling and Pre-processing

Raw data required significant processing to be suitable for machine learning:

- **Filtering:** The dataset was filtered to include only Falcon 9 launches, removing Falcon 1 and Falcon Heavy records.
- **Handling Missing Values:** Null values in the `PayloadMass` column were imputed using the mean mass for that specific class.

- **Outcome Classification:** A binary "Class" variable was created. A value of **1** represented successful landings (True Ocean, True RTLS, True ASDS), while **0** represented failures (False Ocean, False RTLS, None, etc.).

3. Operational Insights & Exploratory Data Analysis (EDA)

3.1 Launch Site Analysis

Analysis of the launch sites revealed distinct operational profiles:

- **KSC LC-39A (Kennedy Space Center):** Identified as the "workhorse" of the program, this site handles the highest volume of successful commercial launches (approx. 42%). It is the primary site for heavy-lift missions.
- **CCAFS SLC-40:** This site's data reflects the "trial and error" phase of the program. It shows a cluster of failures between flight numbers 1 and 20, corresponding to early developmental testing.
- **VAFB SLC-4E:** Located on the West Coast, this site is used exclusively for Polar orbits and demonstrates a very high success rate with almost no recorded failures.
-

3.2 Payload and Orbit Dynamics

Visualizations utilizing scatter plots and bar charts provided counter-intuitive insights regarding payload mass:

- **Heavier $\neq$ Riskier:** Contrary to the assumption that heavier payloads increase risk, payloads exceeding 10,000 kg (typically Starlink missions) showed a near-perfect success record.
- **Mid-Range Variability:** The highest variability in landing success was observed in the mid-range payload category (2,000 kg – 6,000 kg).
- **Orbital Trends:**
    - **LEO & ISS:** These missions represent the program's "bread and butter," appearing consistently throughout the flight history.
    - **GTO (Geostationary Transfer Orbit):** A surge in GTO missions in later flight numbers marks SpaceX's aggressive expansion into the commercial satellite market.
    - **Polar Orbits:** These appear only in later flight numbers, marking a capability milestone achieved after the activation of the VAFB launch site.

3.3 Reliability Growth

A temporal analysis of success rates confirms the program's maturity. Early volatility (2013-2015) visually represents the experimental phase. However, following the introduction of the "Full Thrust" (FT) booster version in 2016, the success trend trajectory moved sharply upward, stabilizing near 100% in recent years.

3.4 Geospatial Findings

Interactive mapping with Folium highlighted critical logistical constraints. All launch sites are located on coastlines for safety reasons. Furthermore, sites are situated in immediate proximity to railway lines, which is a logistical necessity for transporting massive rocket boosters that cannot travel on standard roads.

4. Predictive Analysis & Conclusion

4.1 Machine Learning Model Development

To predict landing outcomes, four classification models were developed and evaluated: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).

- **Preprocessing:** Data was standardized using StandardScaler to ensure that large numerical values (like Payload Mass) did not overpower binary features during training.
- **Optimization:** Each model underwent hyperparameter tuning using GridSearchCV with 10-fold cross-validation to maximize accuracy and prevent overfitting.

4.2 Model Performance and Selection

Remarkably, all four optimized models achieved an identical test set accuracy of **83.33%**.

- **Best Model Selection:** In the event of a performance tie, **Logistic Regression** was selected as the optimal model. This decision was based on Occam's Razor; Logistic Regression is computationally more efficient, easier to interpret, and less prone to overfitting compared to complex models like SVMs or Decision Trees.

4.3 Error Analysis

Confusion matrix analysis revealed a consistent behavior pattern across all models:

- **Strengths:** The models demonstrated 100% recall on successful landings, correctly identifying 12 out of 12 successes in the test set.
- **Weaknesses:** The primary source of error was False Positives. The models occasionally predicted a successful landing for missions that actually failed. This indicates the models are slightly "optimistic" and may underestimate specific rare failure conditions.

4.4 Final Conclusion

The project successfully demonstrated that machine learning can accurately predict SpaceX Falcon 9 landing outcomes. The analysis proves that SpaceX has effectively navigated the "learning curve," transitioning from high-risk experimental flights to a stable, reliable commercial operation.
The ability to predict these outcomes with ~83% accuracy provides significant business value, allowing stakeholders to confidently assess risk and estimate launch costs in the evolving commercial space race.