IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Sajed Karimy
Dec 1, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- **Summary of Methodologies**:
  - Data was collected using SpaceX API and web scraping techniques to ensure comprehensive and clean datasets.
  - Data preprocessing included standardization and encoding for effective analysis.
  - Exploratory data analysis (EDA) and model tuning were performed to gain insights and enhance predictions.
- **Summary of Results**:
  - Decision Tree was identified as the best model with 88.8% accuracy, showcasing superior generalization.
  - Key insights included the impact of payload, orbit type, and experience on mission success rates.
  - Analysis demonstrated SpaceX's iterative improvement in landing success over time.

Introduction

- **Project Background and Context**:
  - SpaceX aims to reduce launch costs by reusing the Falcon 9 first stage. Predicting successful landings is crucial for evaluating cost-effectiveness and mission planning.
  - This analysis supports competitive bidding against SpaceX by identifying patterns in successful landings.
- **Problems You Want to Find Answers**:
  - What factors influence Falcon 9 first-stage landing success?
  - Which machine learning model provides the most accurate predictions of landing outcomes?
  - How do payload mass, orbit type, and launch site impact mission success?

Section 1

Methodology

Methodology

Executive Summary

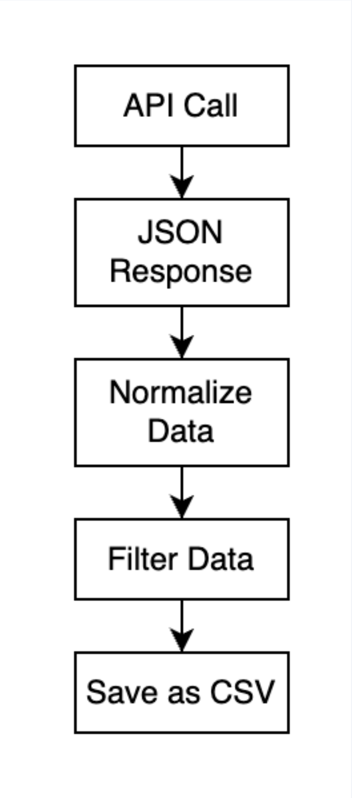
- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- **Sources**: SpaceX API and Wikipedia.
- **Processes**:
    1. API: Fetched structured launch data using REST calls.
    2. Scraping: Extracted historical launch records from Wikipedia tables.
- **Tools**: Python (requests, BeautifulSoup, pandas).

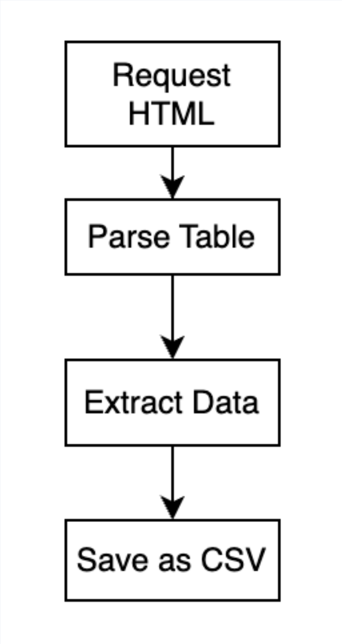
Data Collection – SpaceX API

- Fetched launch data from specexdata.
- Processed JSON response into a pandas DataFrame.
- Extracted relevant details (rocket version, payload, launch site).
- Saved the clean dataset.
- GitHub link



Data Collection - Scraping

- Scraped Wikipedia for launch records using BeautifulSoup.
- Parsed HTML tables, extracted key fields, and cleaned noisy data.
- Created a pandas DataFrame and saved it.
- GitHub Link



Data Wrangling

- **Class Creation**:
  - Successful landing (Class = 1): True ASDS, True RTLS, True Ocean.
  - Unsuccessful landing (Class = 0): False ASDS, False RTLS, False Ocean, None.
- Most launches occurred at **CCAFS SLC 40** (55 launches).
- Frequent orbits: GTO (27), ISS (21), VLEO (14).
- Clean dataset saved as dataset_part_2.csv
- GitHub Link

EDA with Data Visualization

- **Key Charts:**
  1. **Launch Sites Distribution**: Showed frequency of launches from each site (e.g., CCAFS SLC 40 was most common).
  2. **Orbit Distribution**: Analyzed mission orbits (e.g., GTO and ISS were most frequent).
  3. **Payload vs. Success**: Scatterplot of payload mass vs. success rates to identify performance trends.
- GitHub Link

EDA with SQL

- **Queries**
  1. **Launch Sites**: Identified unique launch sites.
  2. **Payload by Customer**: Summed payload masses launched by NASA.
  3. **First Success**: Found the date of the first ground pad success.
  4. **Landing Outcome Analysis**:
     1. Counted successful vs. failed landings.
     2. Ranked outcomes by frequency.
- GitHub Link

Predictive Analysis (Classification)

- **Data Preparation**: Standardized features and split data into training (80%) and testing (20%) sets.
- **Models Accuracies on Test Data**:
  - **Logistic Regression**: **83.3%**
  - **Support Vector Machine (SVM)**: **83.3%**
  - **Decision Tree**: **88.8%**
  - **K-Nearest Neighbors (KNN)**: **83.3%**

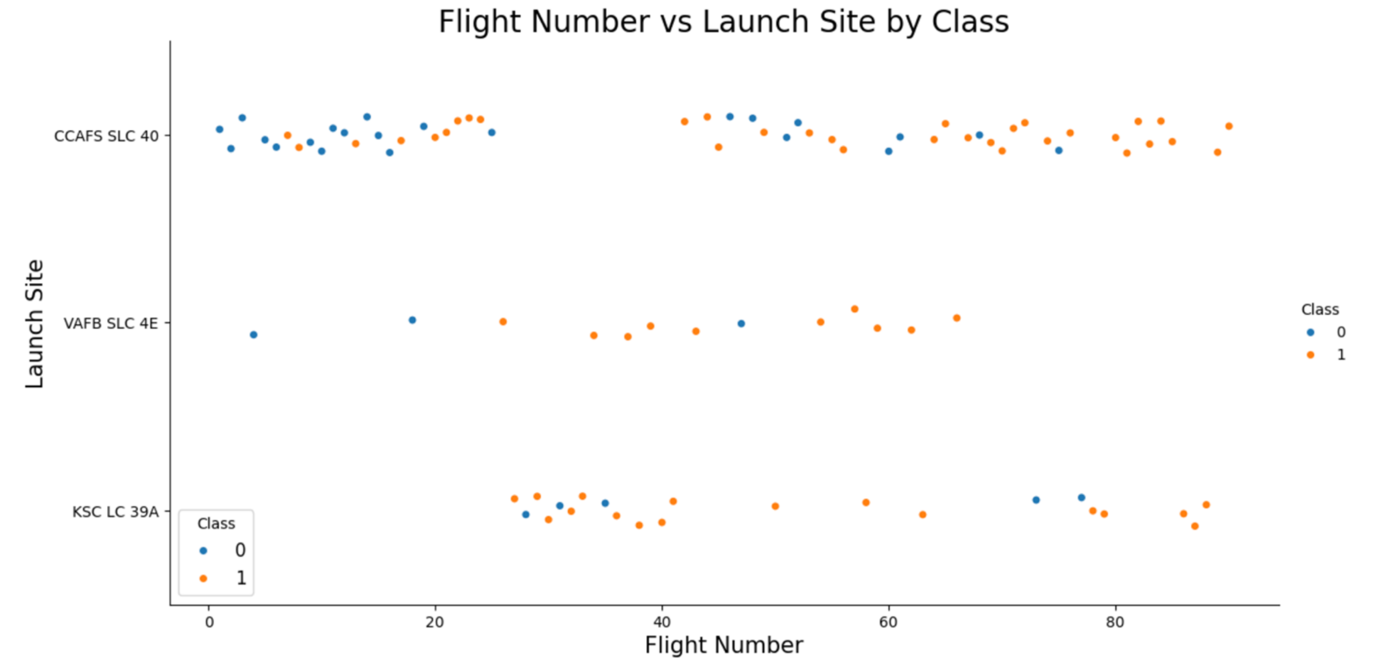
Results

- **Model Evaluation:**
  - **Best Model**: **Decision Tree**:
    - Testing accuracy: **88.8%**.
    - Best balance of performance and reliability.
  - **Other Models**:
    - Logistic Regression, SVM, and KNN showed similar performance with testing accuracy of **83.3%**.
- **Insights:**
  - Decision Tree performed the best due to optimized hyperparameters and effective generalization.
  - Models performed well overall but had some false positives
- GitHub Link

Section 2

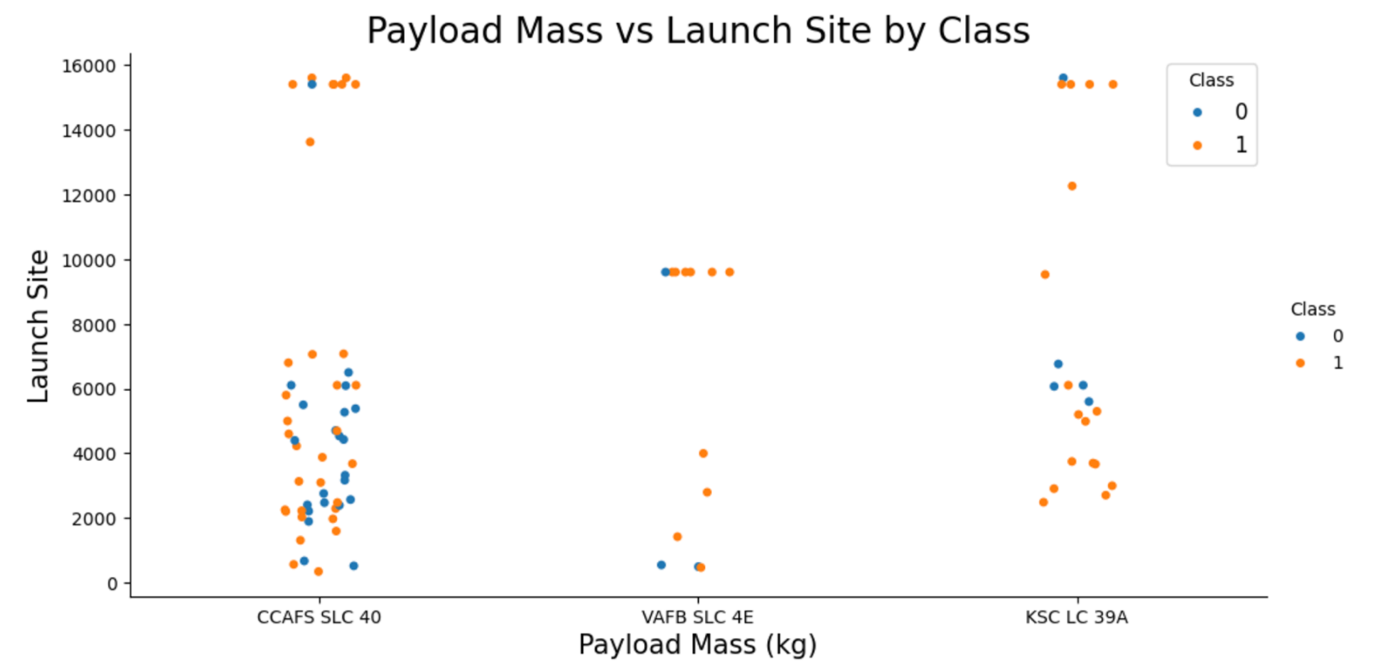
Insights drawn from EDA

Flight Number vs. Launch Site



The chart highlights patterns across sites and time, helping identify successful launch trends based on historical data. The improvement in success rate in later flights demonstrates SpaceX's iterative development approach.

Payload vs. Launch Site

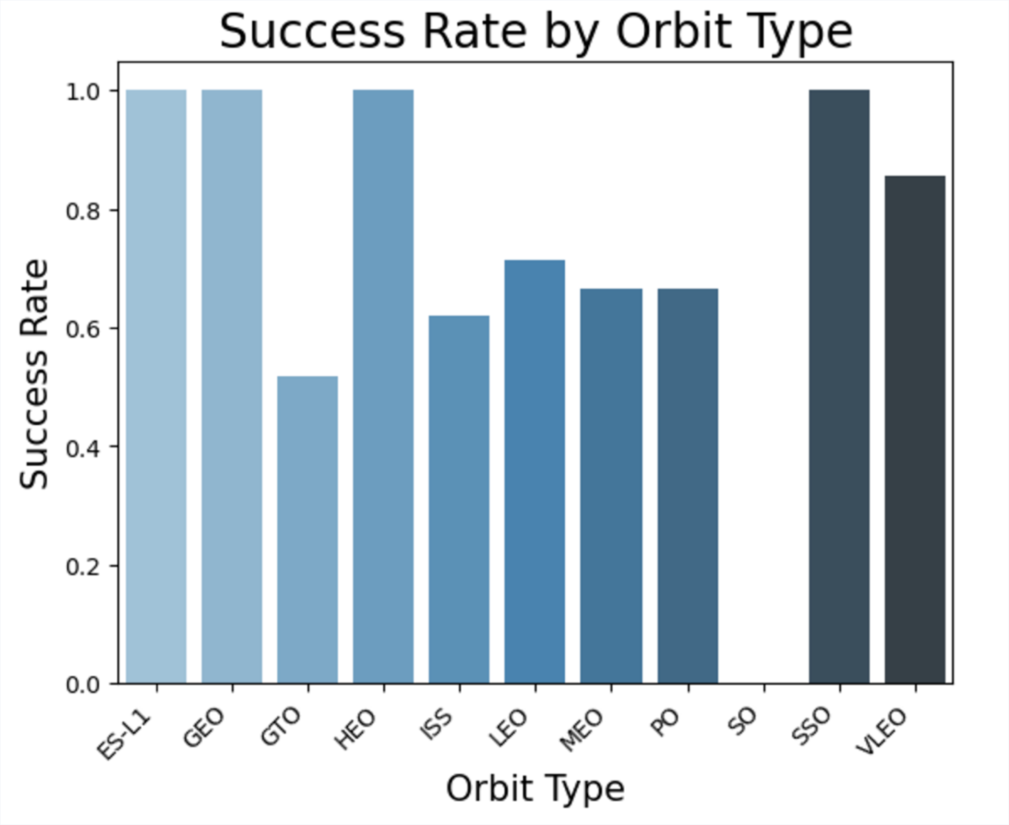


The analysis suggests that payload mass affects the success probability differently across launch sites, with heavier payloads showing a higher failure rate, especially at **VAFB SLC 4E**.

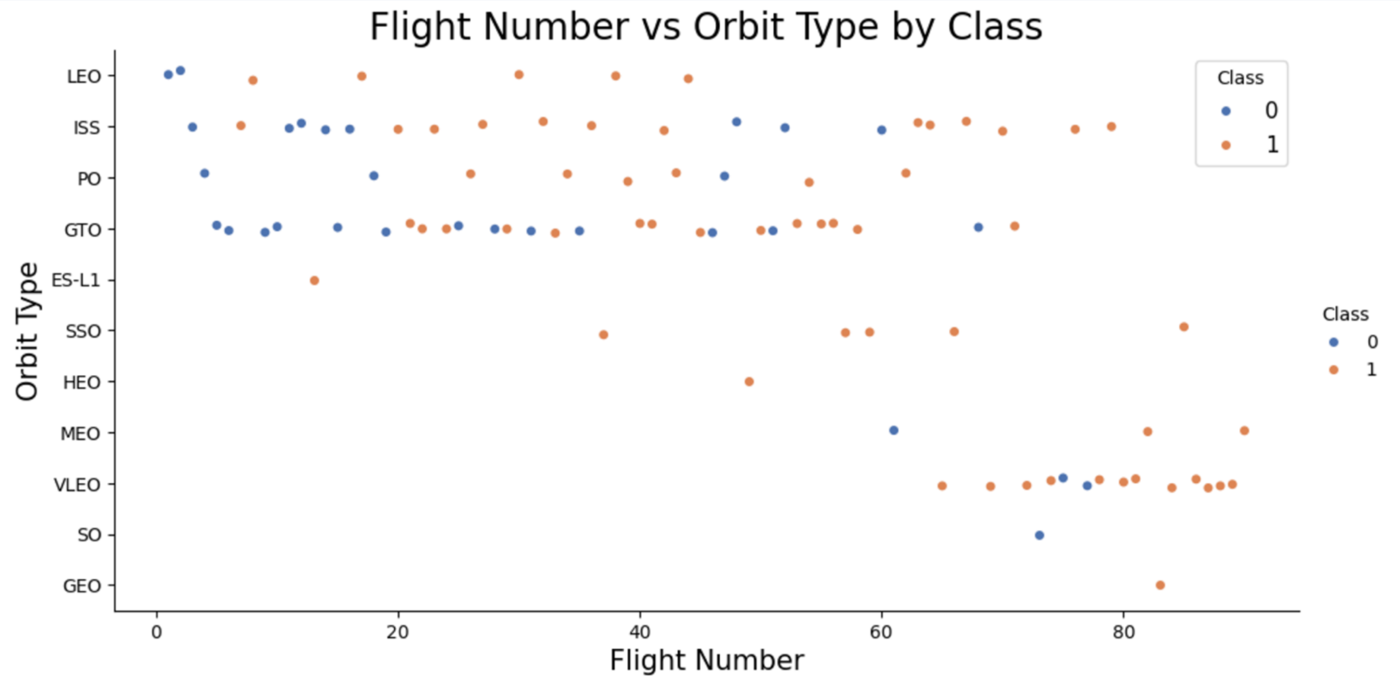
Success Rate vs. Orbit Type

The results suggest mission difficulty increases with orbit altitude or complexity (e.g., GTO and HEO).

This information can guide mission planning and risk assessments based on the target orbit type.

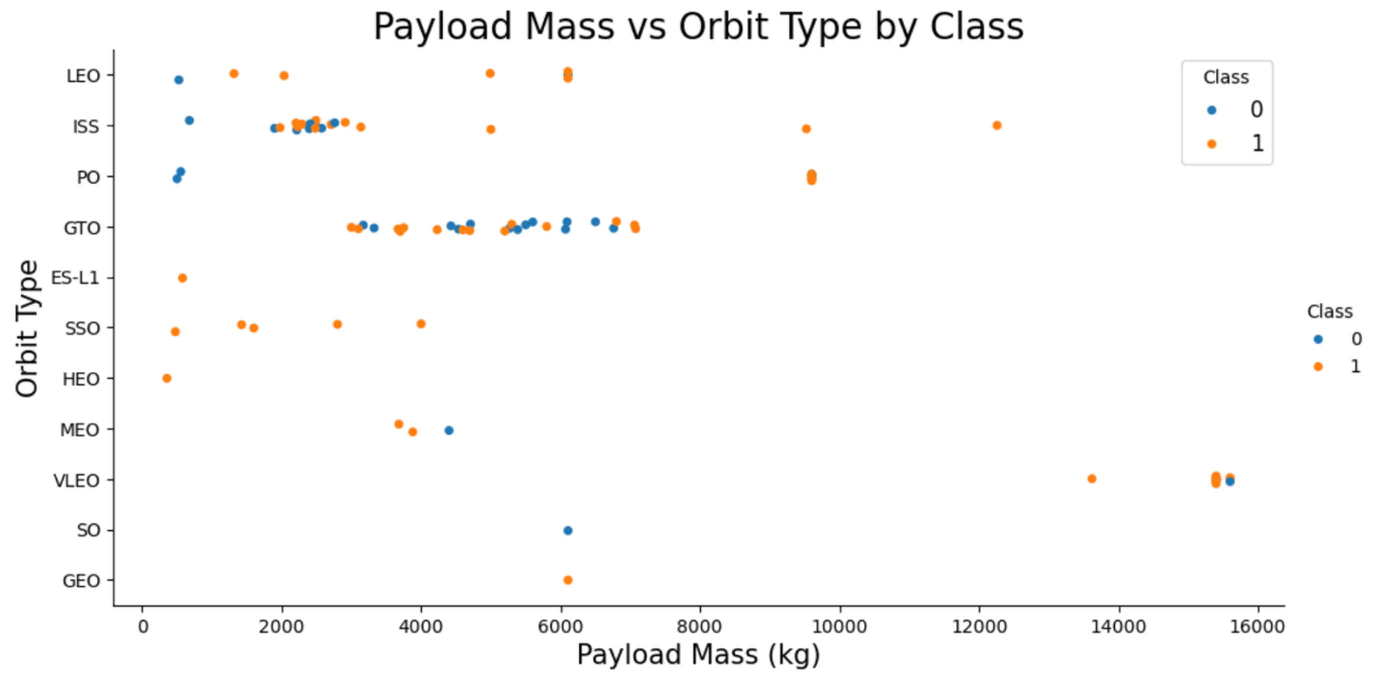


Flight Number vs. Orbit Type



This plot highlights how SpaceX's success rate has improved with experience, especially for challenging orbits like GTO. The data can guide future mission planning to minimize risks and enhance reliability.

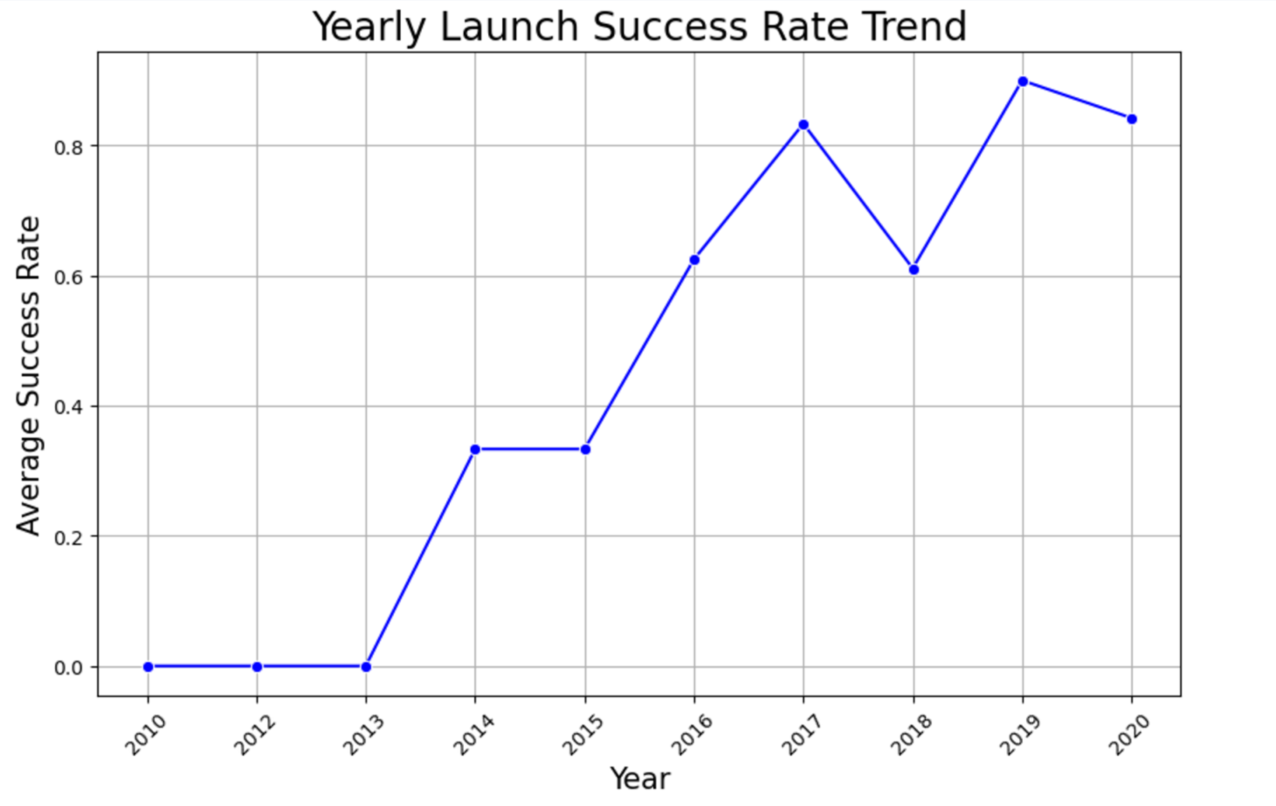
Payload vs. Orbit Type



This plot highlights the relationship between payload mass and mission outcomes for different orbits. The data suggests that higher payloads tend to be associated with higher success rates in certain orbits, while mid-range payloads pose challenges in achieving consistent results.

Launch Success Yearly Trend

The chart highlights SpaceX's steady progress in rocket technology, overcoming early challenges and a brief dip in 2018. The upward trend underscores its reliability and leadership in reusable launch systems.



All Launch Site Names

- Unique Launch Sites Identified:
  1. CCAFS LC-40
  2. VAFB SLC-4E
  3. KSC LC-39A
  4. CCAFS SLC-40
- These are the unique launch sites used by SpaceX for its rocket launches. Each site is strategically chosen based on mission requirements, payload specifications, and orbital destinations.

Launch Site Names Begin with 'CCA'

The table lists the first 5 SpaceX launches from "CCA"-prefix sites (e.g., CCAFS LC-40), highlighting launch dates, payloads, and mission outcomes.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

First Successful Ground Landing Date

The first successful landing on a ground pad occurred on **December 22, 2015**. This milestone marked a significant achievement in SpaceX's reusable rocket technology development.

Total Number of Successful and Failure Mission Outcomes

- **Successful Landings**: 61 (38 general, 14 drone ship, 9 ground pad)
- **Failures**: 12 (5 drone ship, 3 general, 2 parachute, 2 ocean)
- **Other Outcomes**: 28 (5 controlled ocean, 22 no attempt, 1 precluded)

Shows steady success with reusable rockets despite challenges.

Landing_Outcome	outcome_count
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	21
No attempt	1
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

Boosters Carried Maximum Payload

- A large number of booster versions (e.g., F9 v1.0 B0003 to F9 B5 B1058.4) carried the maximum payload mass.
- **Explanation**: This indicates that multiple missions utilized boosters with maximum payload capacity, highlighting consistency across various booster iterations.

2015 Launch Records

- **Result**: In 2015, failures on the drone ship occurred in **January** (F9 v1.1 B1012) and **April** (F9 v1.1 B1015) at **CCAFS LC-40**.
- **Explanation**: These records highlight specific unsuccessful landings during early 2015 for further analysis.

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

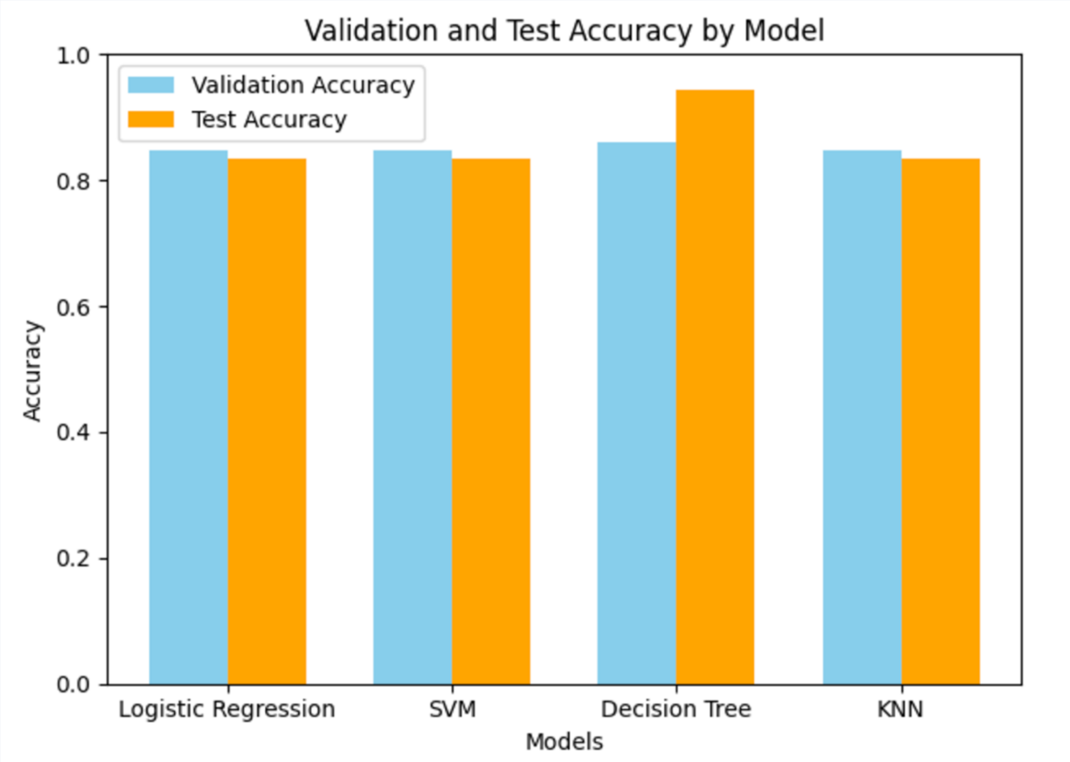
- **Top Outcomes**:
  - **No attempt**: 10 occurrences.
  - **Success (drone ship)**: 5 occurrences.
  - **Failure (drone ship)**: 5 occurrences.
  - **Success (ground pad)**: 3 occurrences.
- **Explanation**: Most missions in this period were "No attempt," while successful landings on drone ships were frequent compared to other outcomes.

Section 3

Predictive Analysis (Classification)

Classification Accuracy

Decision Tree outperforms others with the highest validation and test accuracy, while Logistic Regression, SVM, and KNN show comparable but slightly lower performance.



Conclusions

- **Data Collection and Preparation:** Clean and robust dataset creation was achieved through API calls and web scraping, enabling effective analysis.
- **Top Model Performance:** Decision Tree outperformed with 88.8% accuracy, showcasing superior generalization and hyperparameter optimization.
- **Consistent Success Over Time:** Later flights displayed improved landing success, reflecting SpaceX's iterative development and learning.
- **Payload Influence:** Heavier payloads posed greater risks, impacting mission outcomes significantly.
- **Orbit Complexity:** Higher-altitude orbits (e.g., GTO) presented greater challenges, underscoring the need for precise planning.

Thank you!