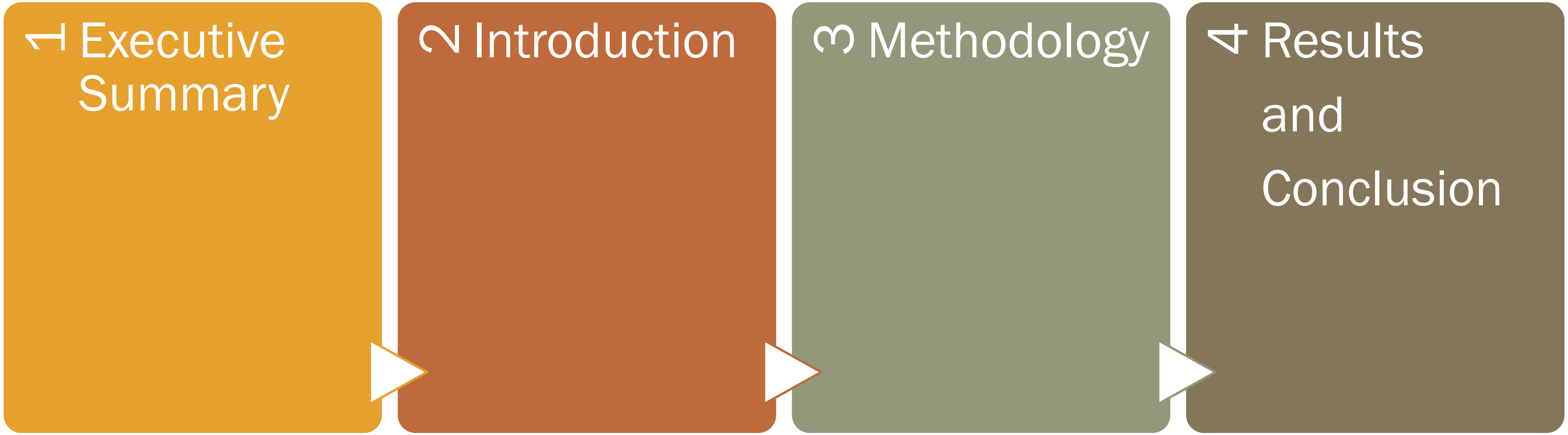
Data Science Capstone Project

HIEU PHAM

https://github.com/hieuthanhpham/Data-Science-Capstone-Project/tree/main

Oct 2024

Outline



1 Executive Summary

2 Introduction

3 Methodology

4 Results and Conclusion

Executive Summary

This project predicts the successful landing of the SpaceX Falcon 9 first stage using machine learning classification algorithms. Key steps include:

- Collected data from SpaceX API and Wikipedia.
- Created a 'class' label for successful landings.
- Explored data using SQL, visualizations, folium maps, and dashboards.
- Converted categorical variables to binary with one-hot encoding and standardized data.
- Used GridSearchCV to optimize machine learning model parameters.
- Built four models: Logistic Regression, SVM, Decision Tree, and KNN.
- Models tended to over-predict success; more data needed for better accuracy.

Introduction

In this capstone, we aim to predict if the Falcon 9 first stage will land successfully.

- SpaceX significantly reduces launch costs by reusing the first stage, offering launches at $62 million compared to competitors at $165 million.
- Knowing if the first stage will land helps estimate the launch cost, useful for competitors bidding against SpaceX.
- Our focus is to predict a successful landing based on features like payload mass, orbit type, and launch site.

Methodology

The overall methodology includes:

1. Data collection, wrangling, and formatting, using:
   - SpaceX API
   - Web scraping

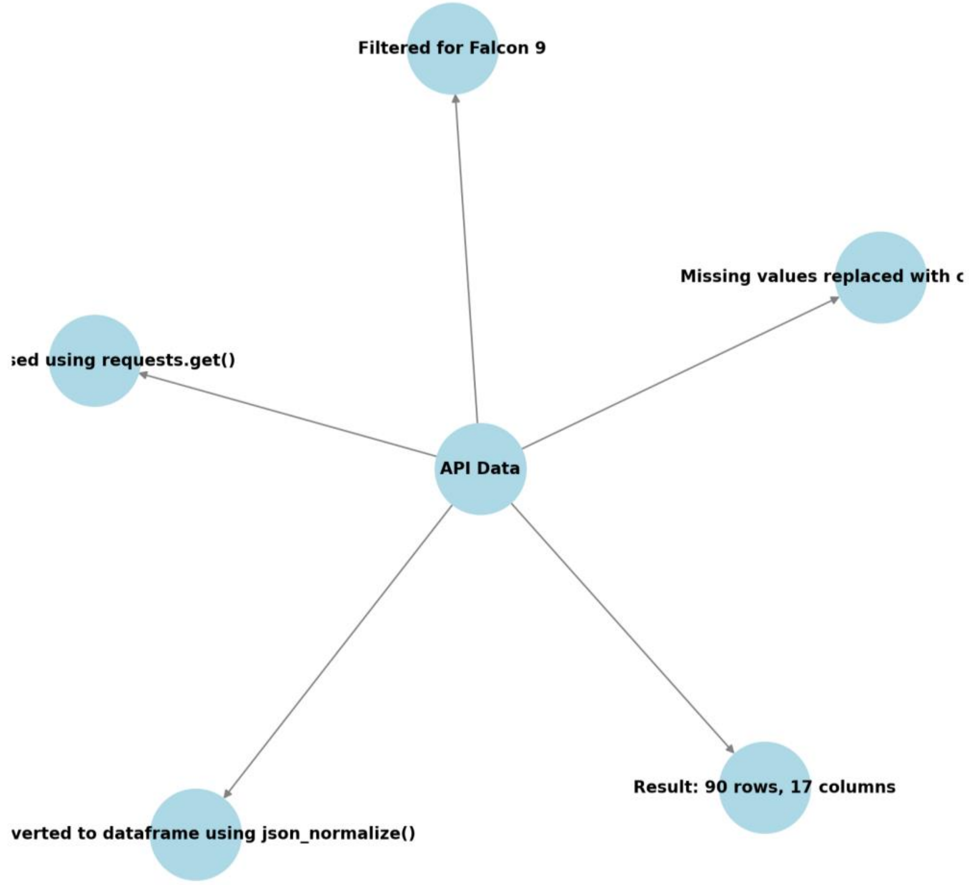
2. Exploratory data analysis (EDA), using:
   - Pandas and NumPy
   - SQL

3. Data visualization, using:
   - Matplotlib and Seaborn
   - Folium
   - Dash

4. Machine learning prediction, using
   - Logistic regression
   - Support vector machine (SVM)
   - Decision tree
   - K-nearest neighbors (KNN)

Key Findings

Data collection using SpaceX API



Filtered for Falcon 9
Missing values replaced with c
sed using requests.get()
API Data
Result: 90 rows, 17 columns
verted to dataframe using json_normalize()

Key Findings

Data Collection with Web Scraping

Libraries or modules used: sys, requests, BeautifulSoup from bs4, re, unicodedata, pandas

•The data is scraped from List of Falcon 9 and Falcon Heavy launches.

- Every missing value in the data is replaced the mean the column that the missing value belongs to.
- We end up with 90 rows or instances and 17 columns or features.

Key Findings

EDA with Pandas and Numpy and SQL

Functions from the Pandas and NumPy libraries such as value_counts() are used to derive basic information about the data collected, which includes:

- The number of launches on each launch site
- The number of occurrence of each orbit
- The number and occurrence of each mission outcome

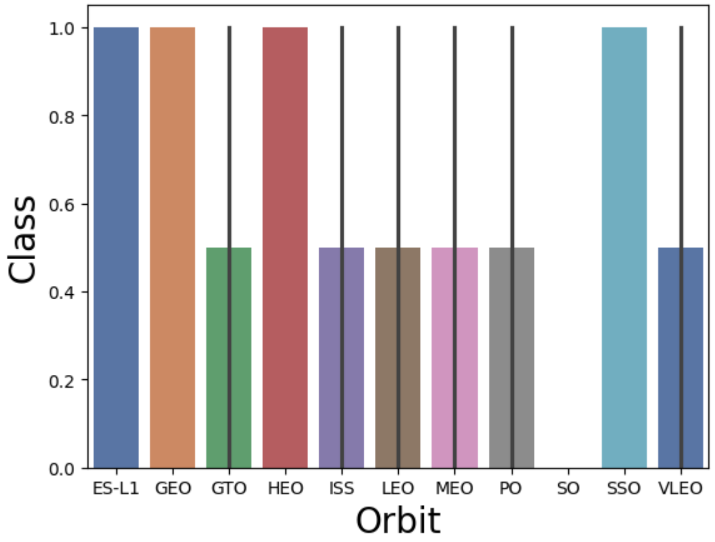
The data is queried using SQL to answer several questions about the data such as:

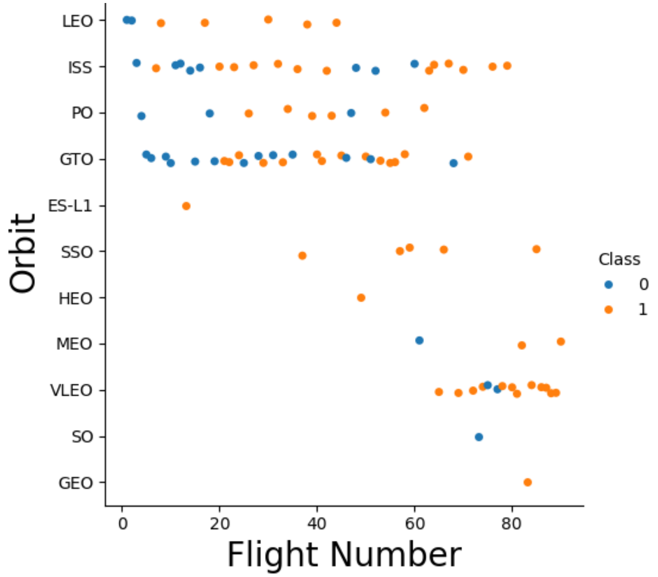
- The names of the unique launch sites in the space mission
- The total payload mass carried by boosters launched by NASA (CRS)
- The average payload mass carried by booster version F9 v1.1
- The SQL statements or functions used include SELECT, DISTINCT, AS, FROM, WHERE, LIMIT, LIKE, SUM(), AVG(), MIN(), BETWEEN, COUNT(), and YEAR().

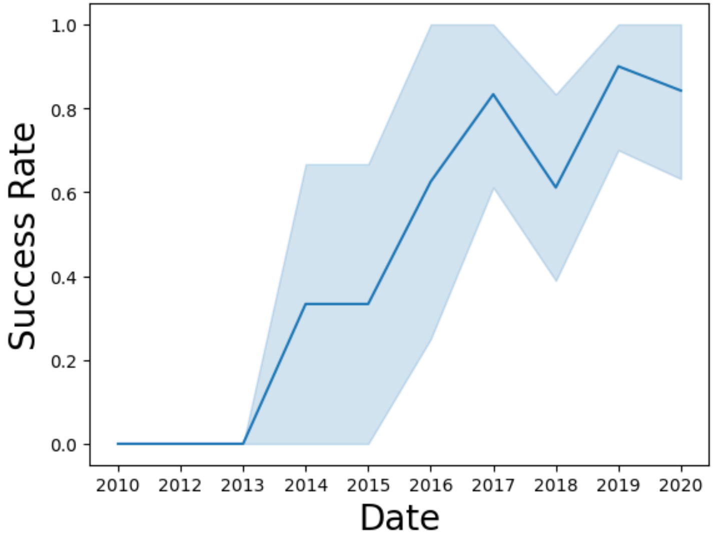
Key Findings

Data Visualization using Matplotlib and Seaborn

- Libraries or modules used: pandas, numpy, matplotlib.pyplot, seaborn
- Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.



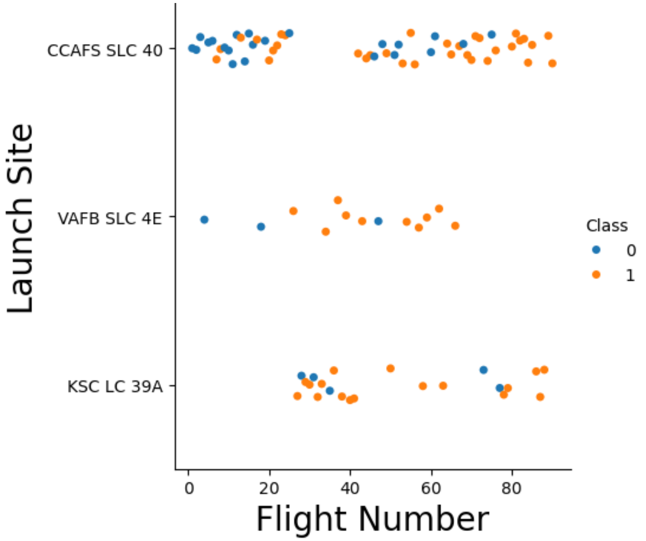


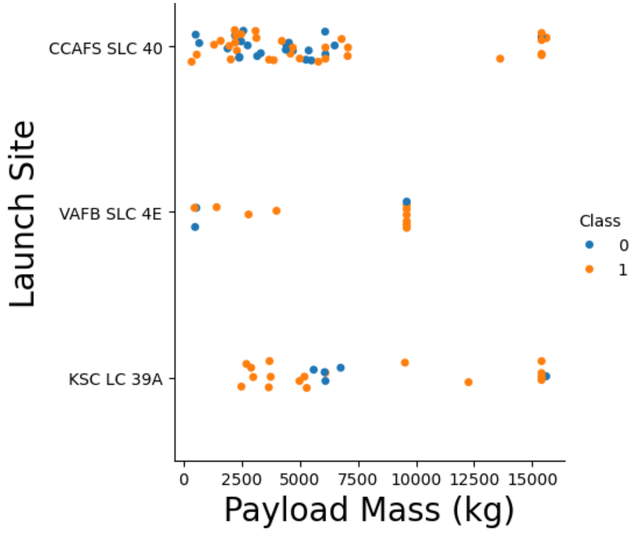


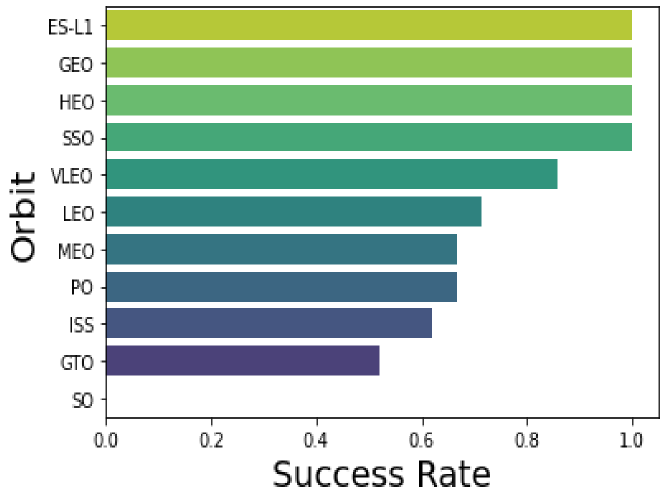
Key Findings

Data Visualization using Matplotlib and Seaborn

- The plots and charts are used to understand more about the relationships between several features, such as:
  - The relationship between flight number and launch site
  - The relationship between payload mass and launch site
  - The relationship between success rate and orbit type

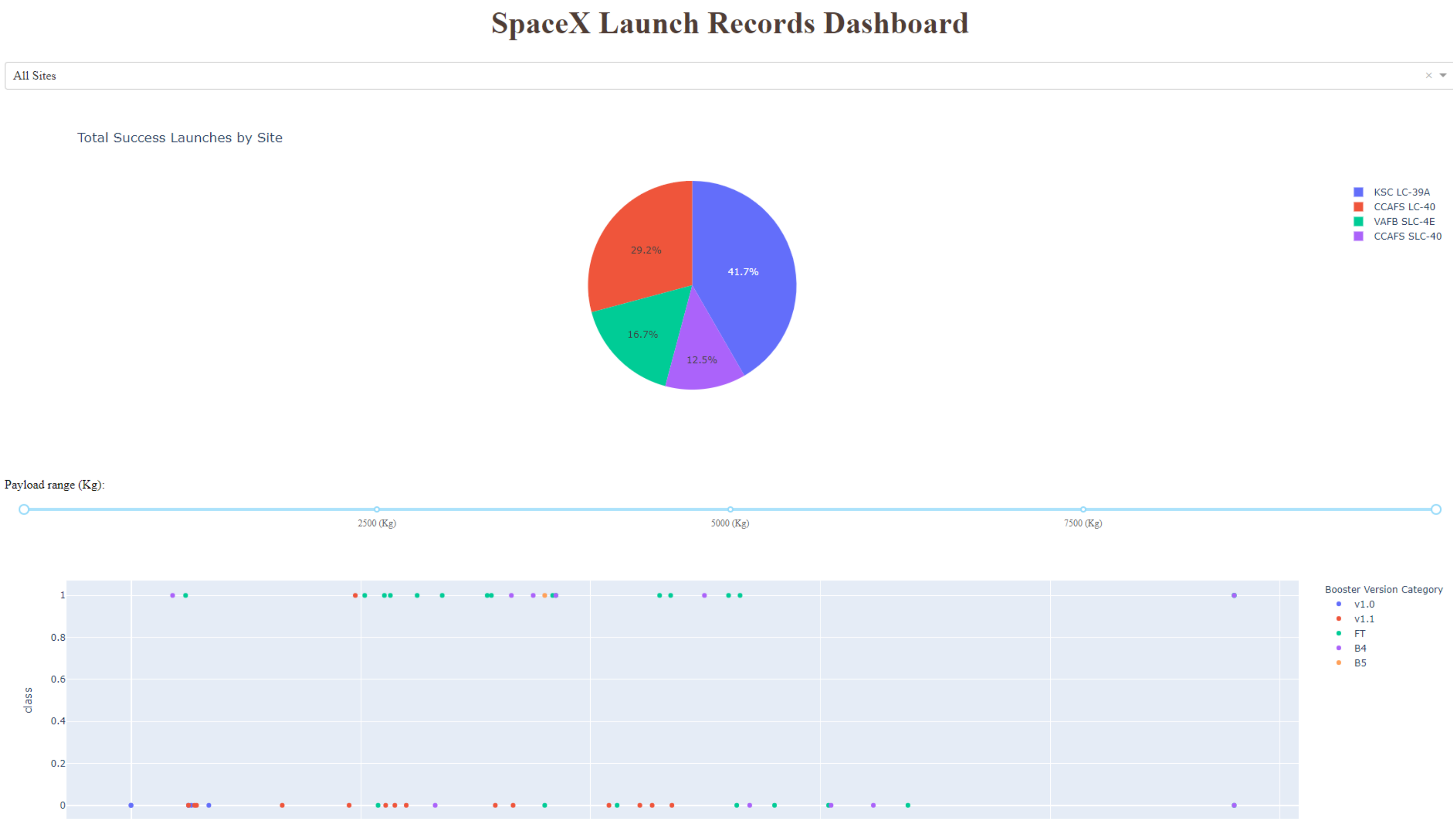






Results

Plotly dashboard

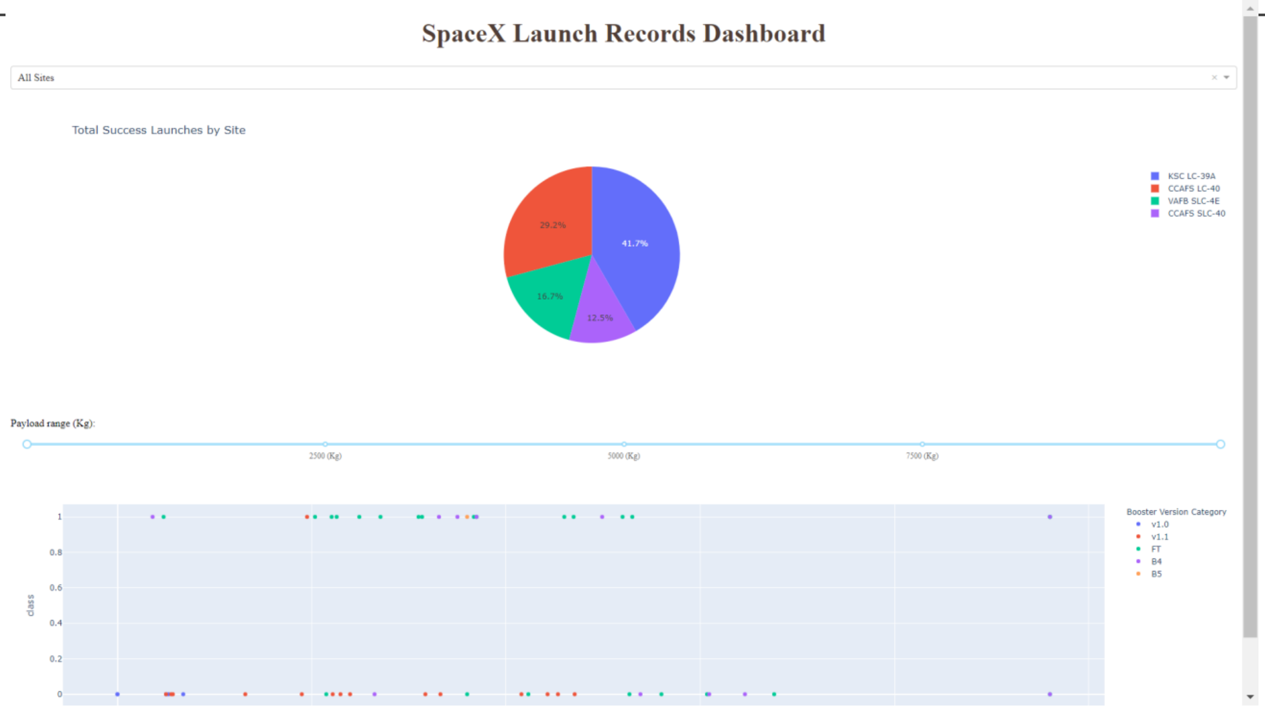


SpaceX Launch Records Dashboard

All Sites

Results

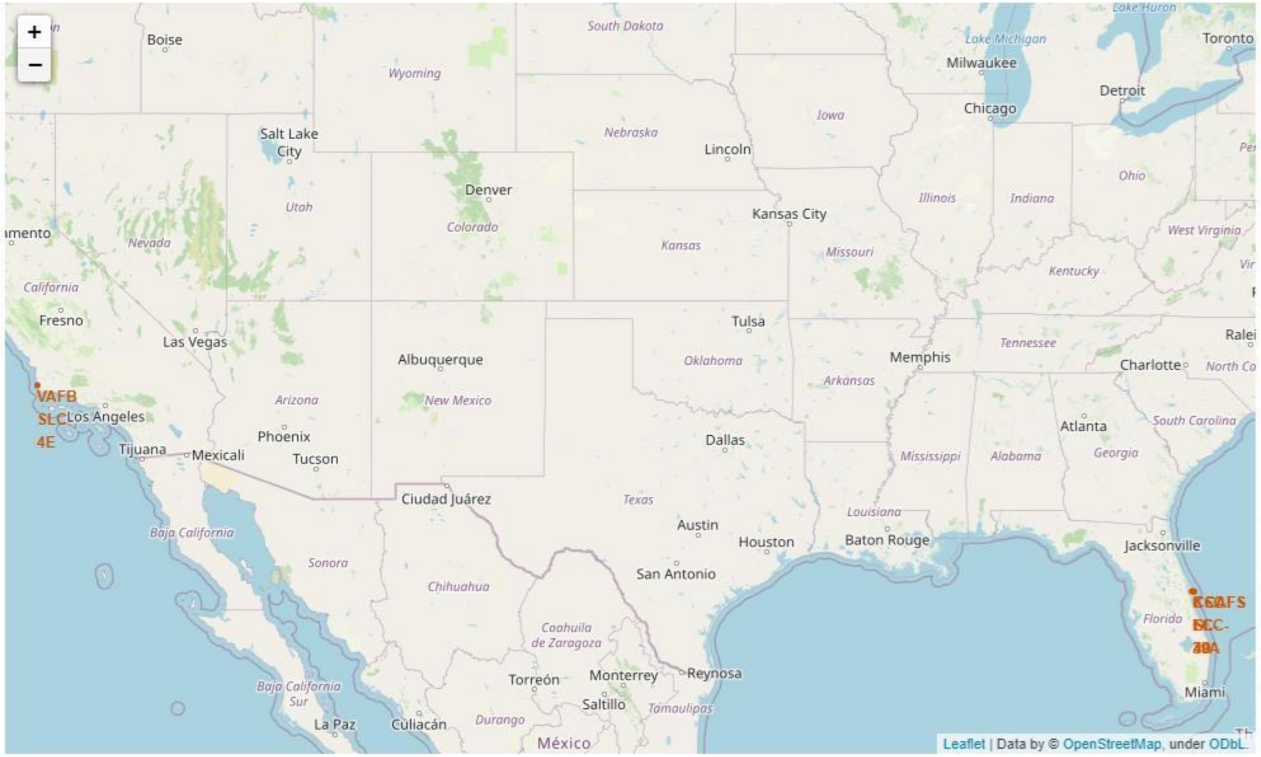
Plotly dashboard

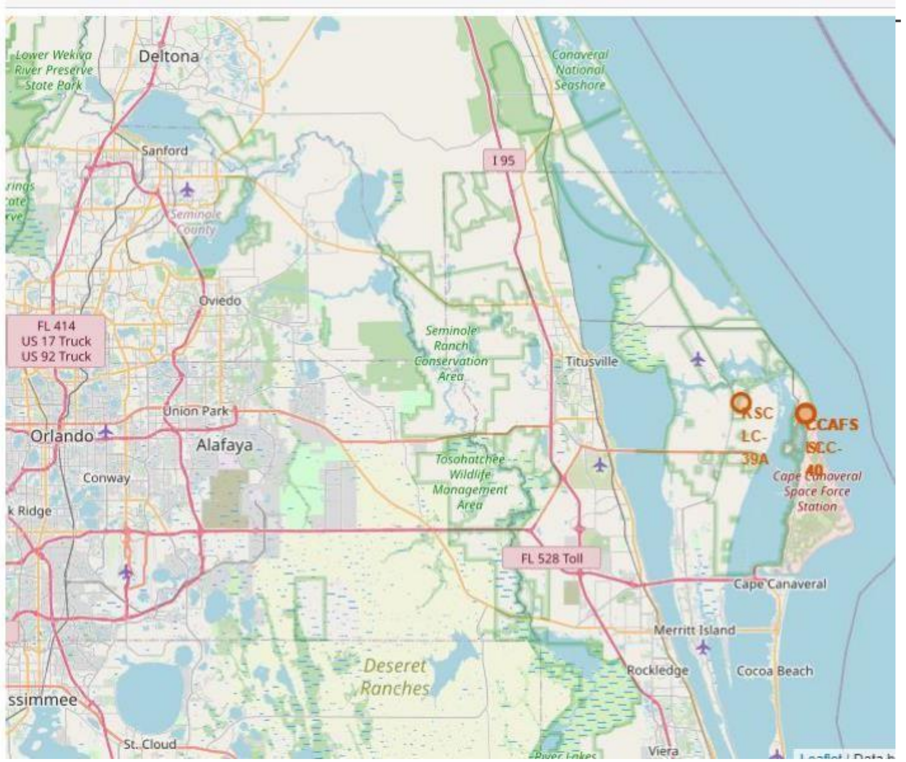


This is a preview of the. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

Interactive Map with Folium

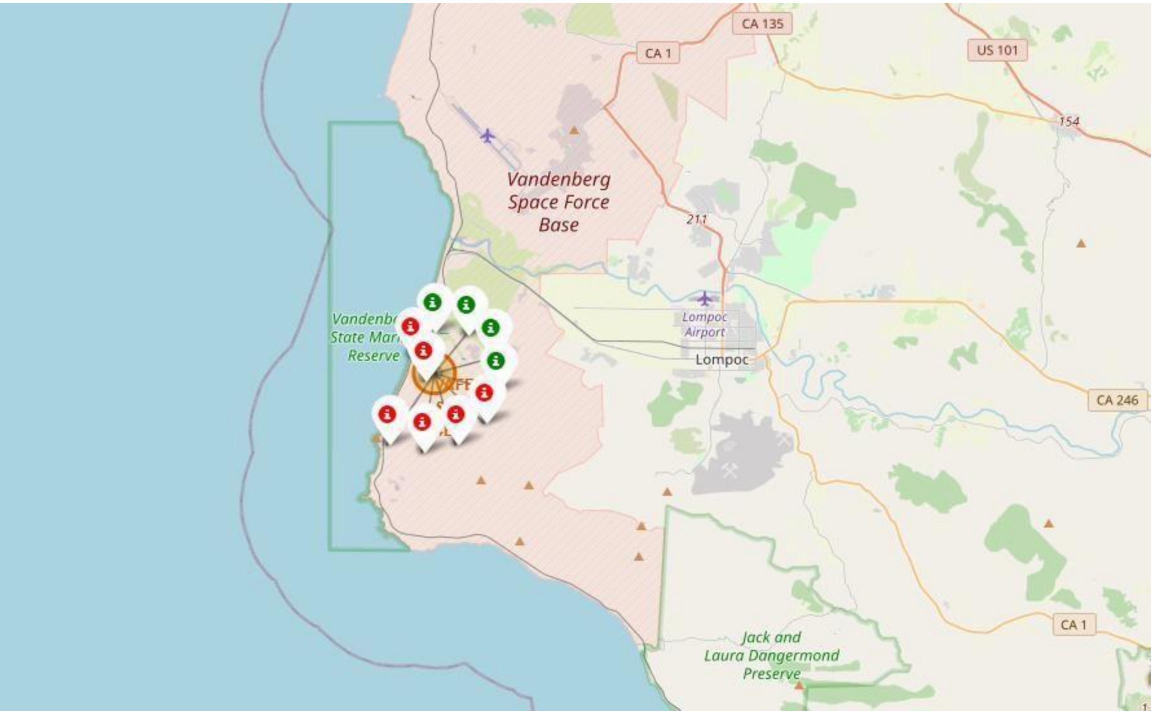
Launch Site Locations





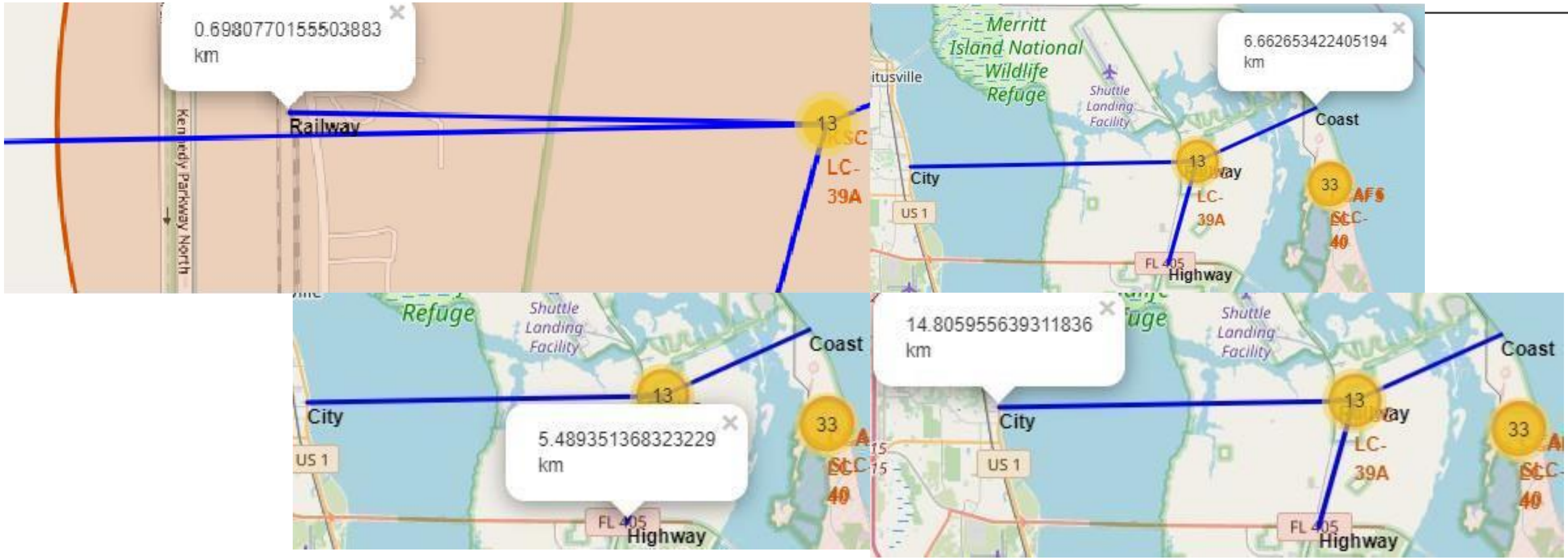
The left map shows U.S. launch sites. The right map highlights the two close Florida sites. All launch sites are near the ocean.

Color-Coded Launch Markers


CA 135
CA 1
US 101
154
Vandenberg Space Force Base
211
Vandenbe State Mari Reserve
Lompoc Airport
Lompoc
CA 246
CA 1
Jack and Laura Dangermond Preserve

Clicking clusters on the Folium map shows successful (green) and failed (red) landings. VAFB SLC-4E has successful and 6 failed landings.

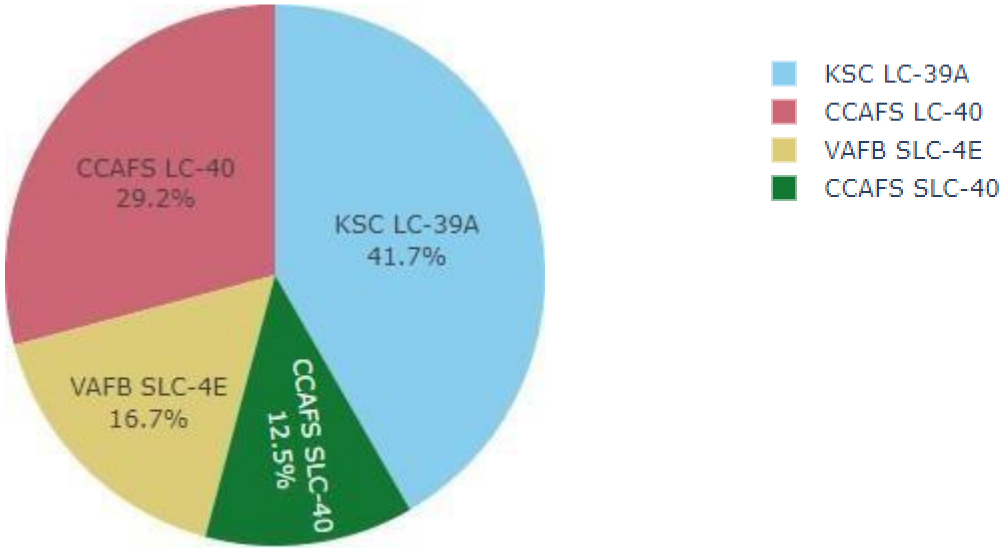
Key Location Proximities



KSC LC-39A shows launch sites near railways for parts, highways for transport, and coasts to ensure failed launches fall safely into the sea, away from cities.

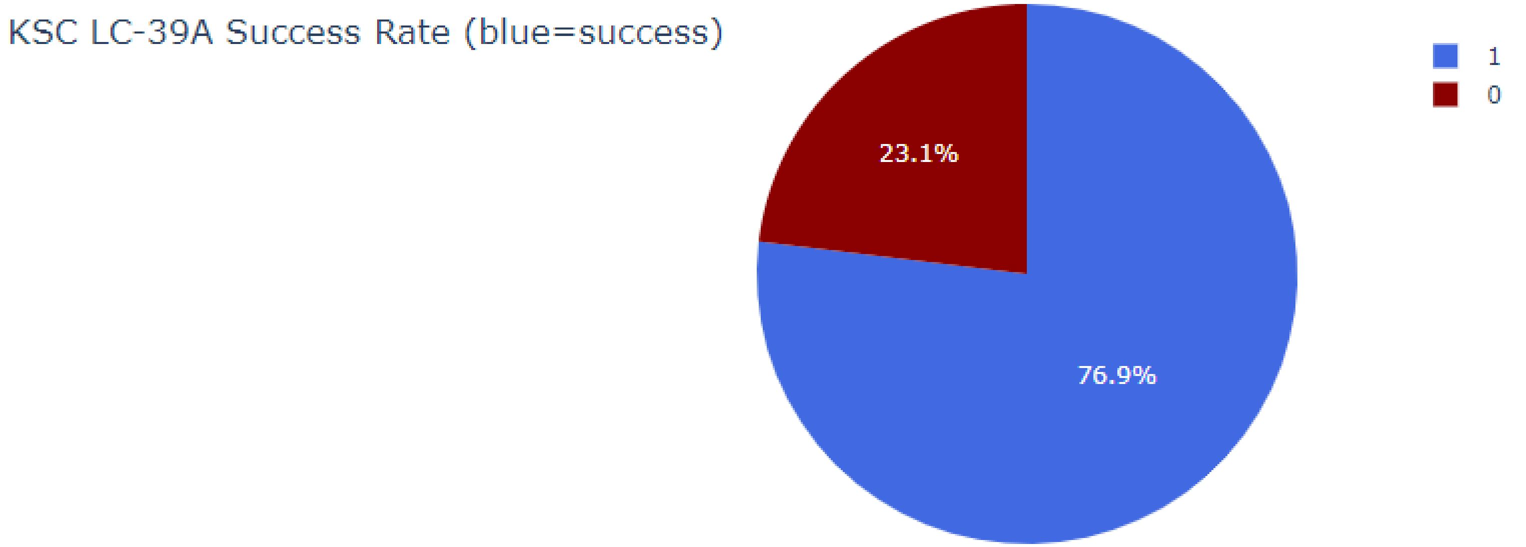
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



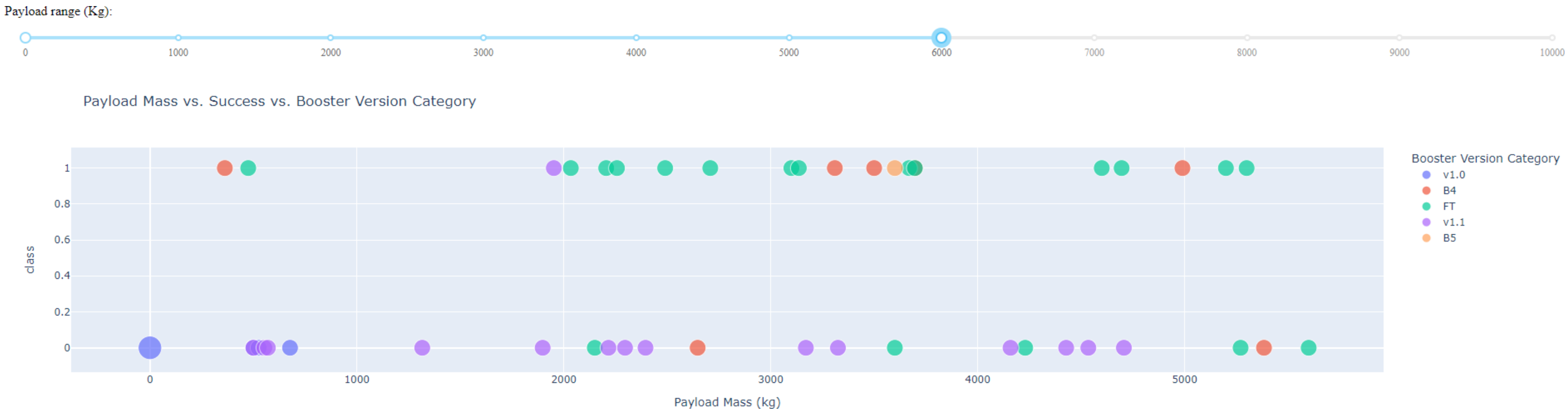
This shows the distribution of successful landings. CCAFS LC-40 (previously CCAFS SLC-40) and KSC have equal landings, mostly before the name change. VAFB has the fewest due to a smaller sample and challenges on the west coast.

Highest Success Rate Launch Site



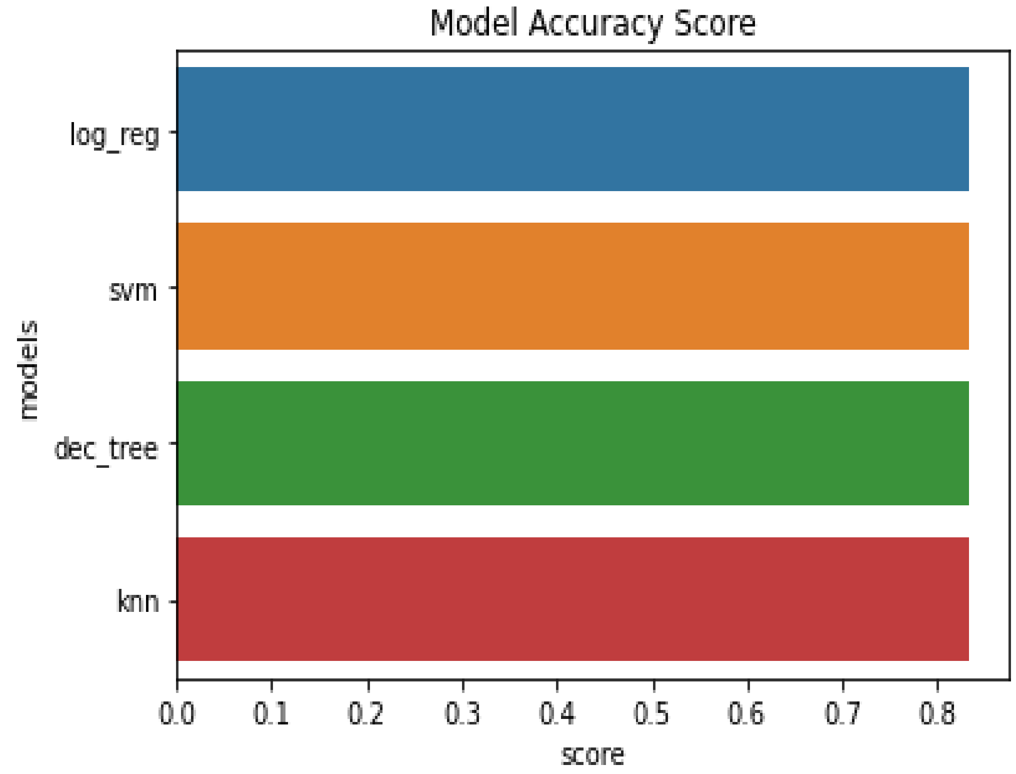
KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

Payload Mass vs. Success vs. Booster Version Category



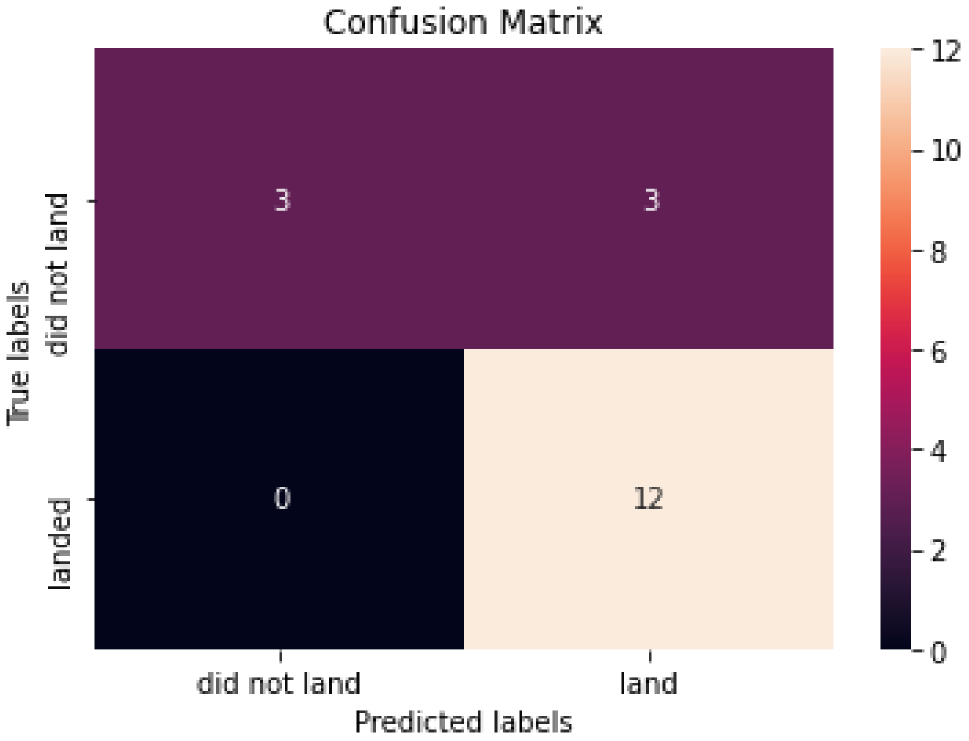
Payload range (Kg):

Classification Accuracy



- All models had similar accuracy (83.33%) on the test set, but the small sample size of 18 may cause high variance, especially in the Decision Tree model.
- More data is needed to identify the best model.

Confusion Matrix



- All models had the same confusion matrix: 12 correct successful landing predictions, 3 correct unsuccessful landing predictions, and 3 false positives.
- The models tend to over-predict successful landings.

Conclusions

Our task was to develop a machine learning model for Space Y to compete with SpaceX.

- The model predicts Stage 1 landing success to save ~$100M, using data from SpaceX API and Wikipedia.
- Data was labeled and stored in a DB2 SQL database, with a dashboard for visualization.
- The model achieved 83% accuracy. SpaceY can use it to predict successful landings before launch, though more data is needed to improve the model.