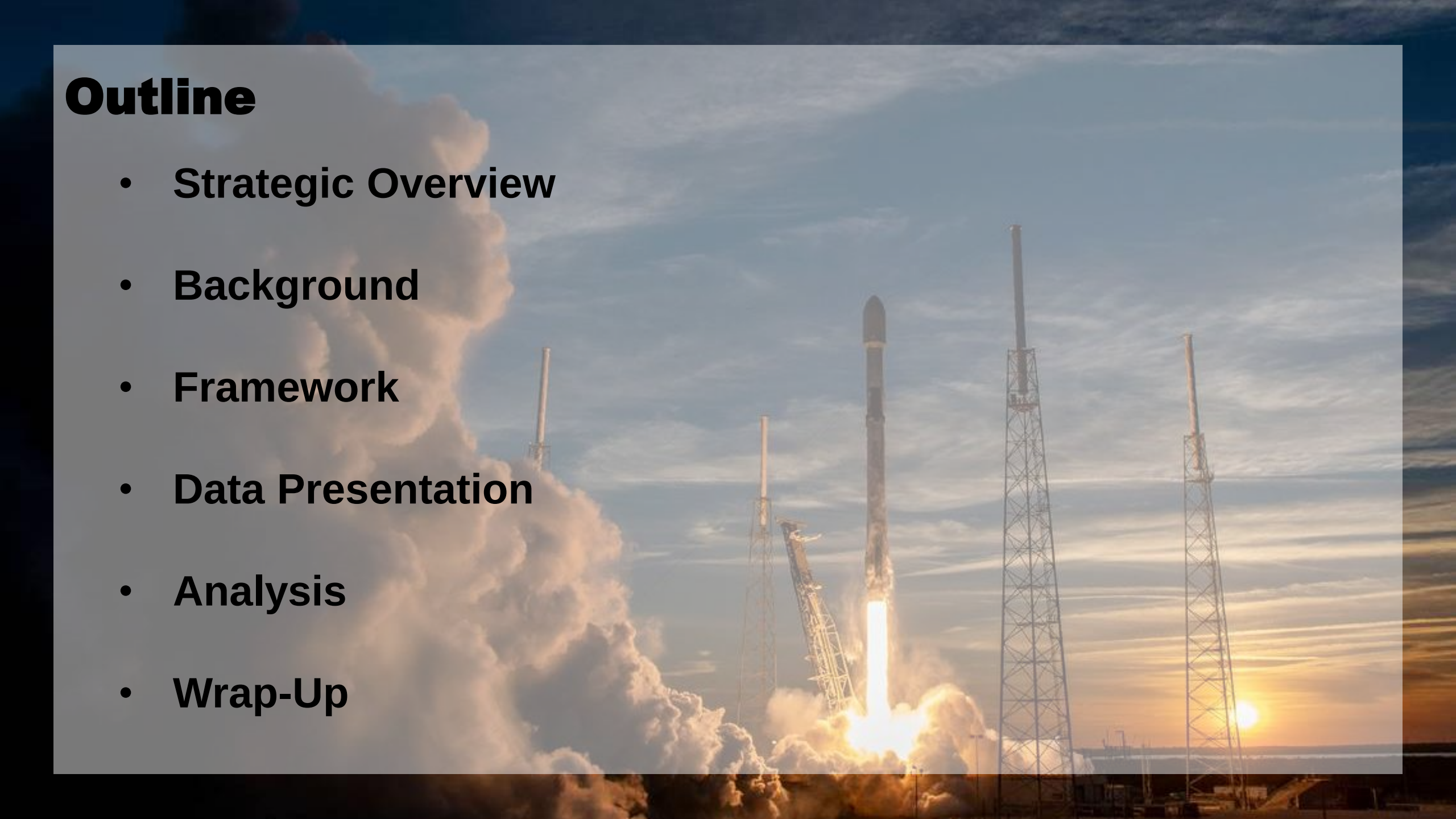
Data Science and Machine Learning Capstone Project

Alberto López García | August, 2024

Outline

- Strategic Overview
- Background
- Framework
- Data Presentation
- Analysis
- Wrap-Up



Strategic Overview

In this comprehensive project, we aim to forecast the successful landing of SpaceX's Falcon 9 first stage using various classification techniques in machine learning. The project is structured around several key phases:

- Gathering, cleaning, and organizing data
- Conducting initial data investigations
- Developing interactive visual representations
- Implementing predictive models through machine learning

Our visualizations reveal specific launch characteristics that are linked to either successful or unsuccessful outcomes. Furthermore, our findings suggest that the decision tree algorithm stands out as particularly effective for predicting the landing outcomes of the Falcon 9's first stage.

Background

In our capstone project, the focus is to assess the probability of successful recoveries of SpaceX's Falcon 9 rocket's first stage. Notably, SpaceX offers these launches at a significantly reduced cost of 62 million dollars, in contrast to competitors who may charge over 165 million dollars. This cost efficiency largely depends on the reusability of the rocket's first stage. Determining the likelihood of its successful landing is thus crucial for evaluating the financial feasibility of launches, especially when potential competitors consider placing bids against SpaceX.

It should also be recognized that some landings that do not succeed are pre-planned, such as controlled descents into the ocean. Our main investigative question explores whether we can predict the success of the first stage landing based on various factors, including the mass of the payload, the type of orbit, and the launch location.

Approach

The methodology for this project is comprehensive and multi-faceted, encompassing several stages:

1. Data Acquisition and Preparation: This initial phase involves:

- Utilizing the SpaceX API for primary data retrieval.
- Implementing web scraping techniques to gather additional data.
- Performing data wrangling and formatting to ensure usability.

2. Exploratory Data Analysis (EDA): Key tools and technologies used include:

- Pandas and NumPy for data manipulation and analysis.
- SQL for querying and managing datasets.

Approach

3. Data Visualization: We employ various tools to create insightful visual representations:

- Matplotlib and Seaborn for static graphing.
- Folium for geospatial mapping.
- Dash for building interactive web-based visualizations.

4. Predictive Modeling: We apply several machine learning algorithms to predict outcomes:

- Logistic Regression to estimate probabilities.
- Support Vector Machine (SVM) for classification tasks.
- Decision Tree to model decisions and their possible consequences.
- K-Nearest Neighbors (KNN) to classify based on feature similarity.

Framework

1. Data Acquisition and Preparation

For this project, we utilized the SpaceX API, accessible at https://api.spacexdata.com/v4/rockets/, to gather comprehensive data on various SpaceX rocket launches. To hone our focus on the Falcon 9 model, we applied filters to the dataset to isolate this specific type of launch. In addressing the issue of missing data within our dataset, each missing value was imputed using the mean of its respective column, ensuring a complete and usable dataset for analysis. After these adjustments, our final dataset comprised 90 instances (rows) and 17 different attributes (columns) relevant to our study. The dataset's structure is illustrated below, where you can see the first few rows showcasing the organized data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Framework

1. Data Acquisition and Preparation

For supplementing our dataset, we performed web scraping on the dedicated Wikipedia page listing Falcon 9 and Falcon Heavy launches, available at this https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922. We focused our scraping efforts specifically to extract information relevant to Falcon 9 launches only. After processing the scraped data, we structured it into a dataset containing 121 instances (rows) and 11 distinct attributes (columns). Below is an illustration of how the dataset appears with the initial few rows, providing a snapshot of the collected data:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Framework

1. Data Acquisition and Preparation

As part of the data processing phase, we ensured that the dataset was thoroughly cleaned to eliminate any missing entries. This involved techniques such as filling missing values appropriately and ensuring data completeness. To make the dataset suitable for machine learning algorithms, categorical features were transformed using one-hot encoding. This method converts categorical variables into a form that could be provided to ML algorithms to do a better job in prediction. Furthermore, we introduced an additional column labeled ‘Class’ to the dataset. This column plays a crucial role in our analysis by categorizing each launch as either a failure (0) or a success (1), based on the outcome of the launch. This binary classification helps in simplifying the prediction model. After these enhancements, the processed dataset comprised 90 instances (rows) and expanded to 83 columns (features), due to the one-hot encoding of categorical variables. This expanded feature set provides a comprehensive base for our predictive modeling.

Framework

2. Exploratory Data Analysis (EDA):

In this project, we utilized the powerful functionalities of the Pandas and NumPy libraries to analyze and summarize the data collected. Specific tasks accomplished using these libraries included:

- **Calculating the number of launches at each site**: This involved grouping the data by the launch site and counting the number of instances for each group.
- **Determining the frequency of each orbit type**: By grouping the data by orbit type, we were able to calculate how frequently each orbit was used.
- **Assessing the distribution of mission outcomes**: This required counting the occurrences of each type of mission outcome (e.g., success, failure).

Framework

2. Exploratory Data Analysis (EDA):

In addition to using Pandas and NumPy for preliminary data exploration, SQL was employed to perform more targeted queries on the dataset, allowing for deeper insights into specific aspects of the launch data. Queries executed included:

- **Identifying unique launch sites**: This query helped to ascertain the various launch sites utilized in the space missions.
- **Calculating the total payload mass for NASA (CRS) launches**: We aggregated the payload masses of launches specifically for NASA under the Commercial Resupply Services contract.
- **Determining the average payload mass for Falcon 9 version 1.1 boosters**: This involved calculating the mean payload mass carried by this specific version of the Falcon 9 booster, highlighting the capabilities of different booster versions.

Framework

3. Data Visualization

To enhance our understanding of the SpaceX Falcon 9 launch data, we employed visualization techniques using the Matplotlib and Seaborn libraries, which are powerful tools for creating static and interactive visualizations in Python.

Matplotlib and Seaborn Visualizations:

- **Scatter plots**: These were utilized to examine relationships between continuous variables such as the payload mass and the flight number for each launch site, providing insights into operational patterns.
- **Bar charts**: Effective for categorical data, bar charts helped in visualizing the frequency of launches from different sites and the success rates associated with different orbit types.
- **Line charts**: Used to track changes over time, such as the evolution of launch success rates or payload capacities over successive flights.

Framework

3. Data Visualization

Folium for Interactive Mapping:

Folium, a Python library that facilitates the creation of interactive maps using Leaflet.js, was instrumental in geospatial analysis of the launch sites. Key implementations included:

- **Marking launch sites**: Each launch site was pinpointed on the map, providing a geographical overview of SpaceX operations.
- **Distinguishing mission outcomes**: We differentiated between successful and failed launches at each site using distinct markers, allowing for an immediate visual distinction of performance across different locations.
- **Proximity markers**: Additional markers were used to show important proximities such as the nearest city, railway stations, or highways to each launch site. This helps in assessing logistical advantages or challenges faced by each launch site.

Framework

3. Data Visualization

For this project, we leveraged Dash, a Python web framework built on top of Plotly.js, to develop an interactive web application that allows users to explore and interact with the SpaceX Falcon 9 launch data more dynamically. Dash enables the creation of HTML components using Python, and it's especially useful for building data visualization interfaces.

Key Features of the Dash Application:

- **Interactive Components**: We incorporated a dropdown menu and a range slider within the dashboard. The dropdown menu allows users to select specific launch sites for detailed examination, while the range slider enables filtering of the data based on criteria such as date range, payload mass, or flight number.

Framework

3. Data Visualization

- **Data Visualization Tools**:

  - *Pie Chart*: This chart provides a visual breakdown of the total successful launches from each selected launch site, enabling users to quickly gauge the success rate relative to other sites.
  - *Scatter Plot*: Designed to analyze the relationship between payload mass and mission outcomes, this plot helps identify trends or patterns, such as whether heavier payloads have higher success rates or specific challenges.

Interactive Data Analysis:

- The interactive site, powered by Dash, not only visualizes data but also allows users to manipulate the view according to their specific interests or areas of focus. For instance, users can choose to view data from a particular year, or focus on launches with specific payload criteria, adjusting the visual outputs in real time.

Framework

4. Learning Prediction

The machine learning prediction phase of our project leverages the Scikit-learn library, a robust toolset for implementing various machine learning algorithms. The following steps outline our methodology for developing predictive models to determine the success of SpaceX Falcon 9 first stage landings:

1. **Standardizing the Data**: To ensure that our model handles all features equally, we normalize the dataset. This step helps remove any bias that might be caused by the scales of the variables, making the training process more efficient and effective.

2. **Splitting the Data**: The dataset is divided into two parts: a training set and a test set. Typically, the training set comprises a larger portion of the data, which is used to train the models, while the test set is reserved to evaluate their performance.

Framework

4. Learning Prediction

3. **Creating Machine Learning Models**: We implement several machine learning algorithms to find the most effective one for our prediction task:

   - **Logistic Regression**: A statistical model that predicts the probability of a binary outcome.

   - **Support Vector Machine (SVM):** A powerful classification technique that works well in high-dimensional spaces.

   - **Decision Tree**: A model that uses a tree-like graph of decisions and their possible consequences.

   - **K-Nearest Neighbors (KNN)**: A method that classifies each data point based on how its neighbors are classified.

Framework

4. **Fitting the Models on the Training Set:** Each model is trained using the training dataset to learn from the features.

5. **Hyperparameter Tuning:** This involves experimenting with various settings for the models to determine the best combination of parameters that improves their performance.

6. **Model Evaluation:**

   - ***Accuracy Scores*:** Each model's accuracy is assessed by comparing the predicted results with the actual outcomes in the test set.

   - ***Confusion Matrix*:** This tool helps visualize the performance of an algorithm. It shows the correct and incorrect predictions made by the model across different classes, providing insight into how well the model is predicting each class.

Data Presentation

The results of our analysis of SpaceX Falcon 9 launch data are organized into five key sections, each employing specific tools and techniques for insight extraction:

1. **SQL (EDA with SQL)**: This section demonstrates the exploratory data analysis performed using SQL queries. It includes the extraction of specific information such as unique launch sites, payload masses for different missions, and average payload masses for various booster versions. This helps in understanding the distribution and characteristics of the data at a granular level.

2. **Matplotlib and Seaborn (EDA with Visualization)**: Here, we utilize Matplotlib and Seaborn to create a series of visualizations that illustrate the relationships within the data. Scatter plots, bar charts, and line charts display relationships and trends such as the correlation between flight numbers and launch sites, payload mass versus launch outcomes, and success rates across different orbit types.

Data Presentation

3. **Folium**: This part of the report uses Folium to provide interactive geographical visualizations. Maps highlight the locations of launch sites and distinguish between successful and unsuccessful launches at each site. Additional markers indicate proximity to key infrastructure like cities, railways, and highways, offering insights into logistical advantages or challenges for each site.

4. **Dash**: Dash is used to create an interactive web application that allows users to manipulate the data visualization through user interfaces such as dropdown menus and range sliders. The visual outputs here include pie charts showing the distribution of successful launches by site and scatter plots that explore the relationship between payload mass and launch success.

Data Presentation

5. **Predictive Analysis**: The final section focuses on the application of machine learning algorithms to predict launch outcomes. We detail the process of standardizing data, splitting it into training and testing sets, fitting models like Logistic Regression, SVM, Decision Trees, and KNN, and tuning their hyperparameters. Model performance is evaluated using accuracy scores and confusion matrices, with each model's effectiveness at predicting launch outcomes (class 0 for failure, class 1 for success) thoroughly analyzed.

Data Presentation

1. SQL (EDA with SQL)

- Catalog of all unique launch sites involved in the space missions

Launch_Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- Five specific entries for launch sites that commence with 'CCA'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Data Presentation

1. SQL (EDA with SQL)

- The cumulative payload mass transported by NASA's CRS mission boosters.

Total payload mass by NASA (CRS)
45596

- The mean payload mass transported by the Falcon 9 version 1.1 boosters.

Average payload mass by Booster Version F9 v1.1
2928

- The date of the inaugural successful ground pad landing.

Date of first successful landing outcome in ground pad
2015-12-22

Data Presentation

1. SQL (EDA with SQL)

- Names of boosters that achieved success on a drone ship with payloads between 4000 and 6000 kilograms.

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The aggregate count of successful and failed mission outcomes.

number_of_success_outcomes	number_of_failure_outcomes
100	1

Data Presentation

1. SQL (EDA with SQL)

- The booster versions that have transported the heaviest payloads.

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

Data Presentation

1. SQL (EDA with SQL)

- Failed drone ship landing outcomes in 2015, including booster versions and launch site names.

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

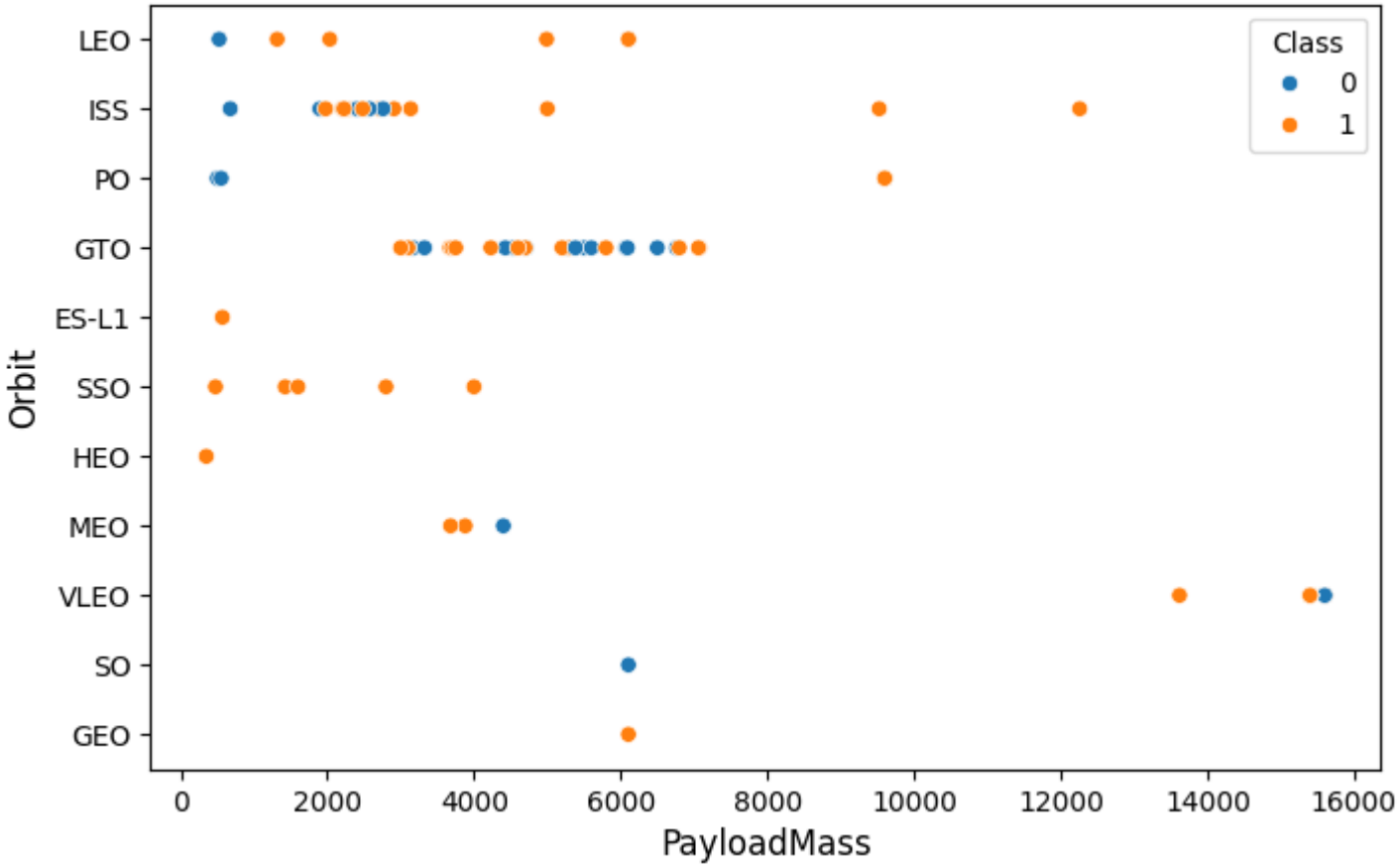
- The count of landing outcomes between 2010-06-04 and 2017-03-20, listed in descending order.

landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Data Presentation

2. Matplotlib and Seaborn

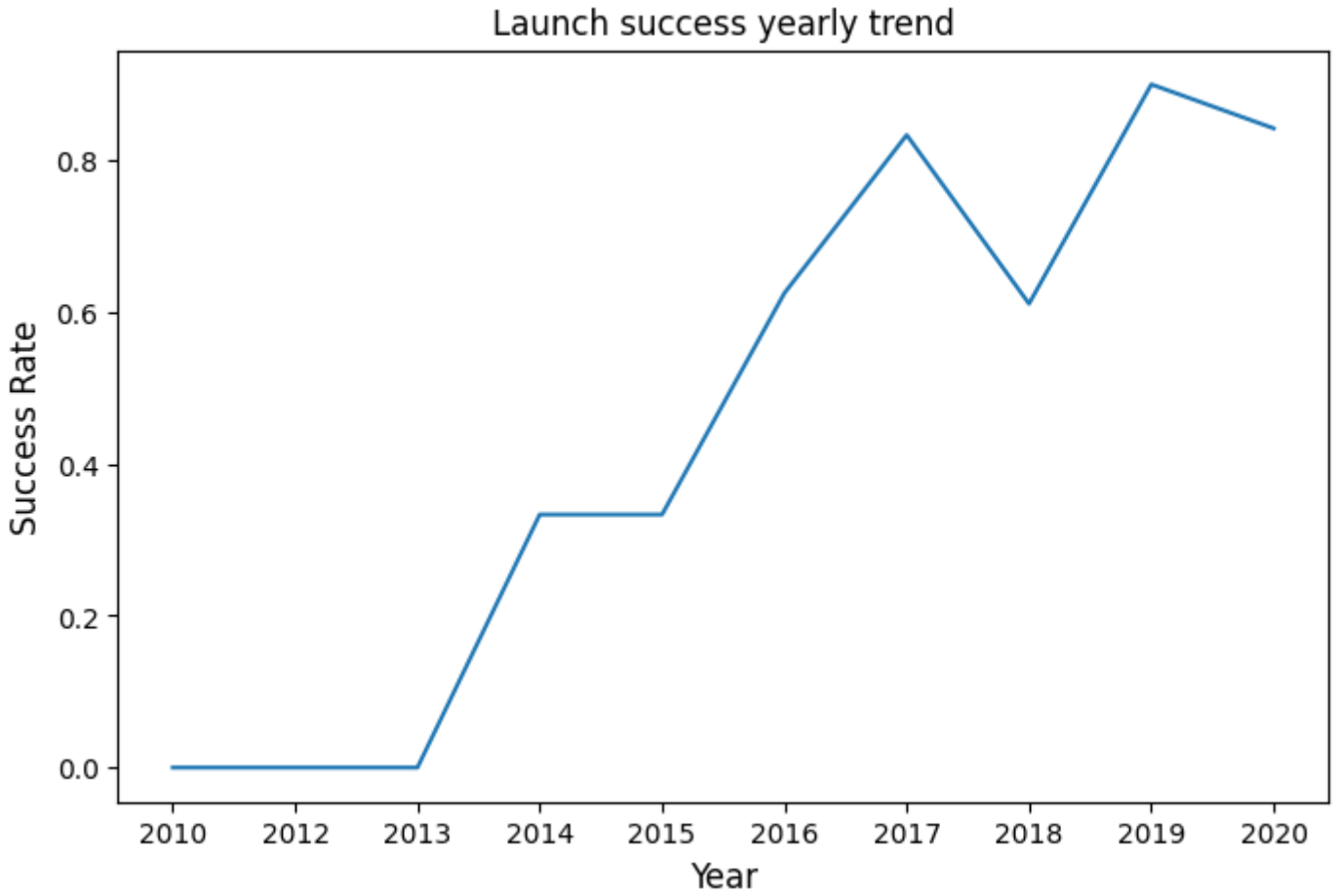
- The correlation between payload mass and orbit type.



Data Presentation

2. Matplotlib and Seaborn

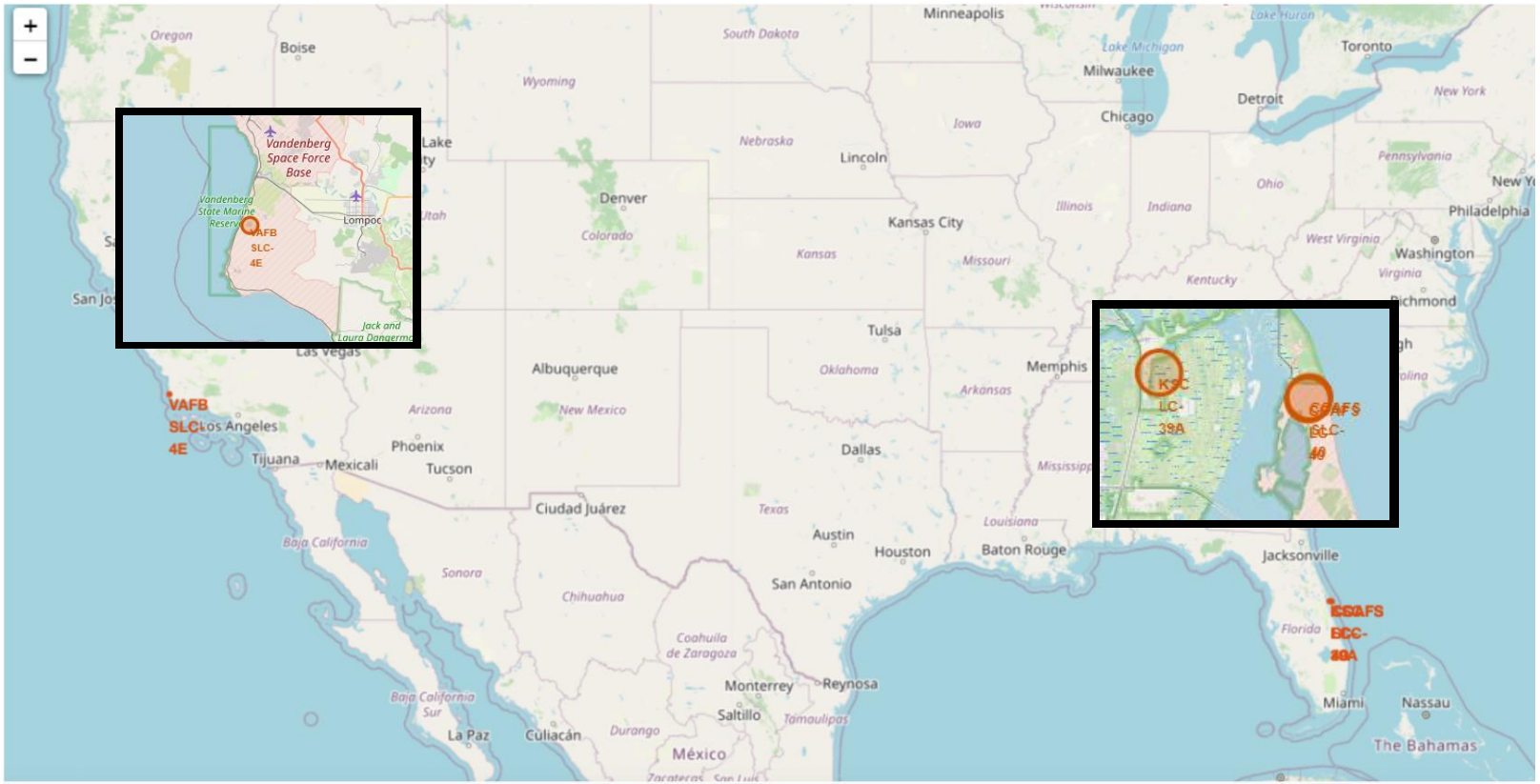
- The annual trend in launch success rates.



Data Presentation

3. Folium

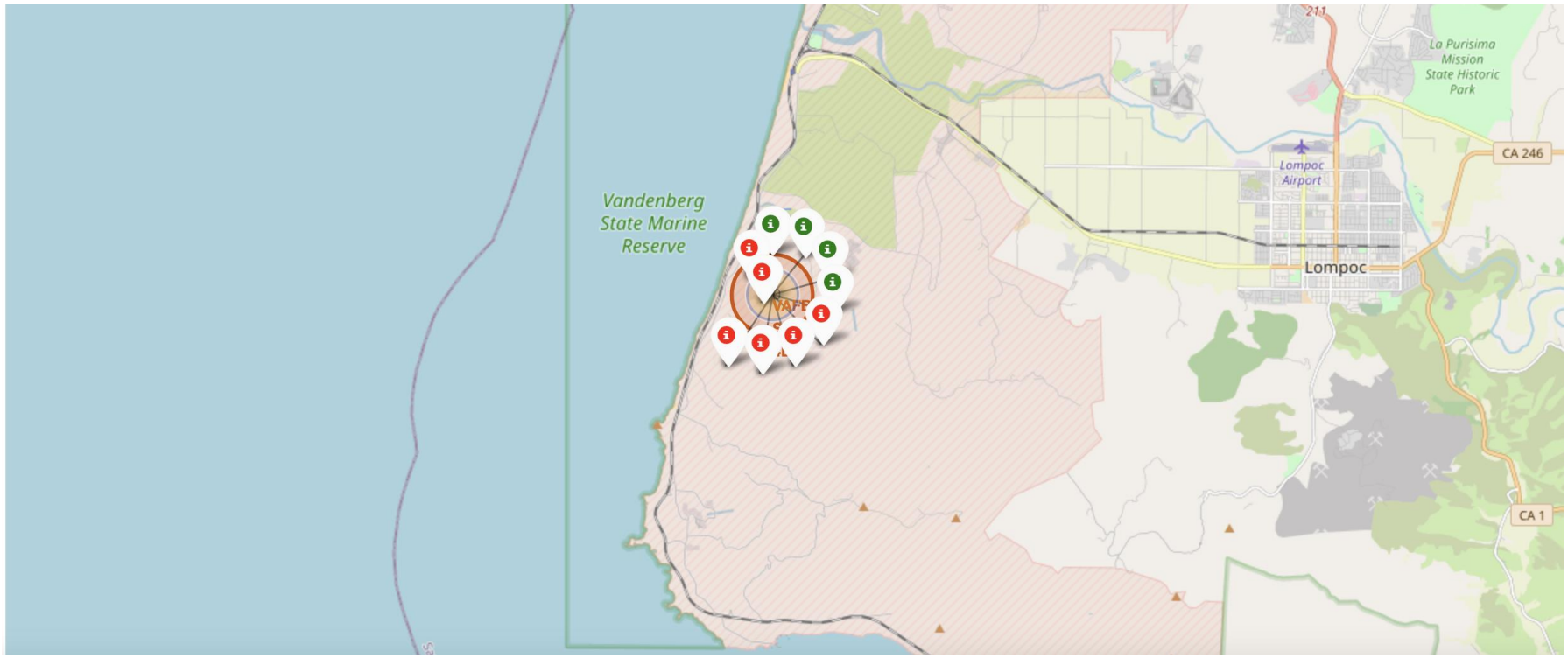
- All launch sites displayed on a map.



Data Presentation

3. Folium

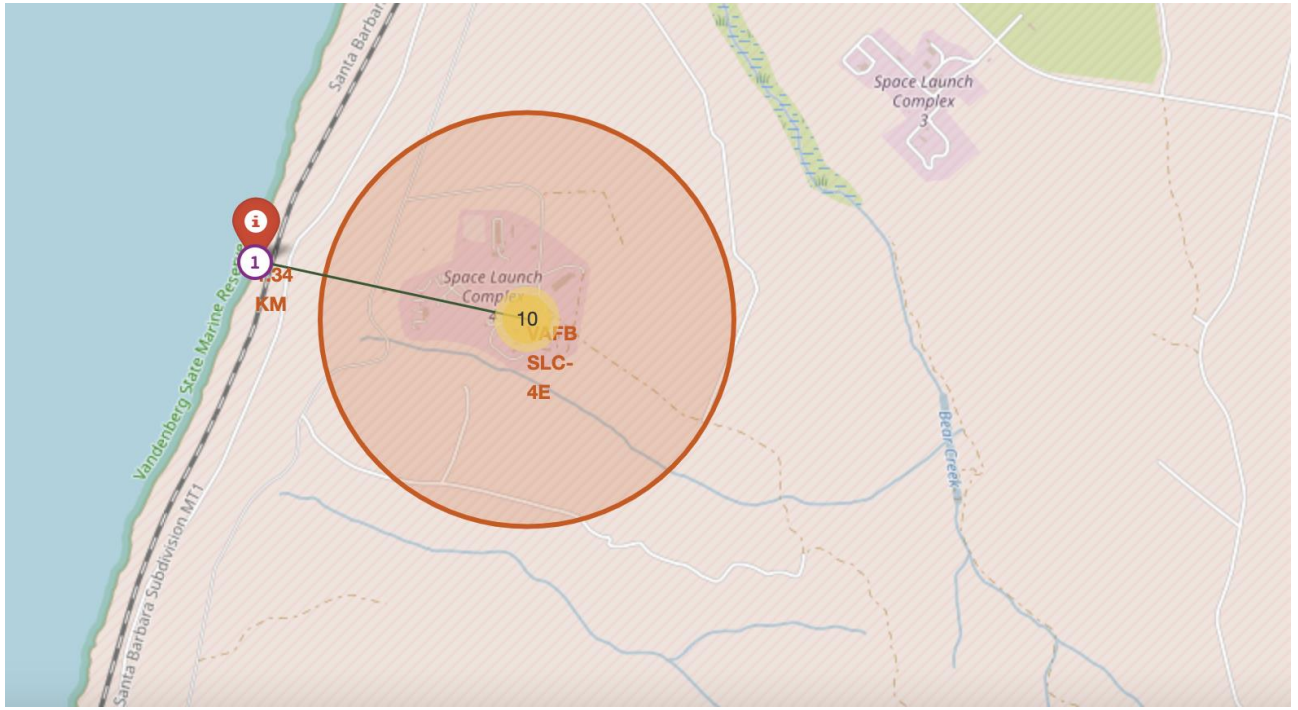
- The successful and failed launches for each site displayed on a map. Zooming in on a launch site reveals green and red tags, where green tags indicate successful launches and red tags indicate failed launches.



Data Presentation

3. Folium

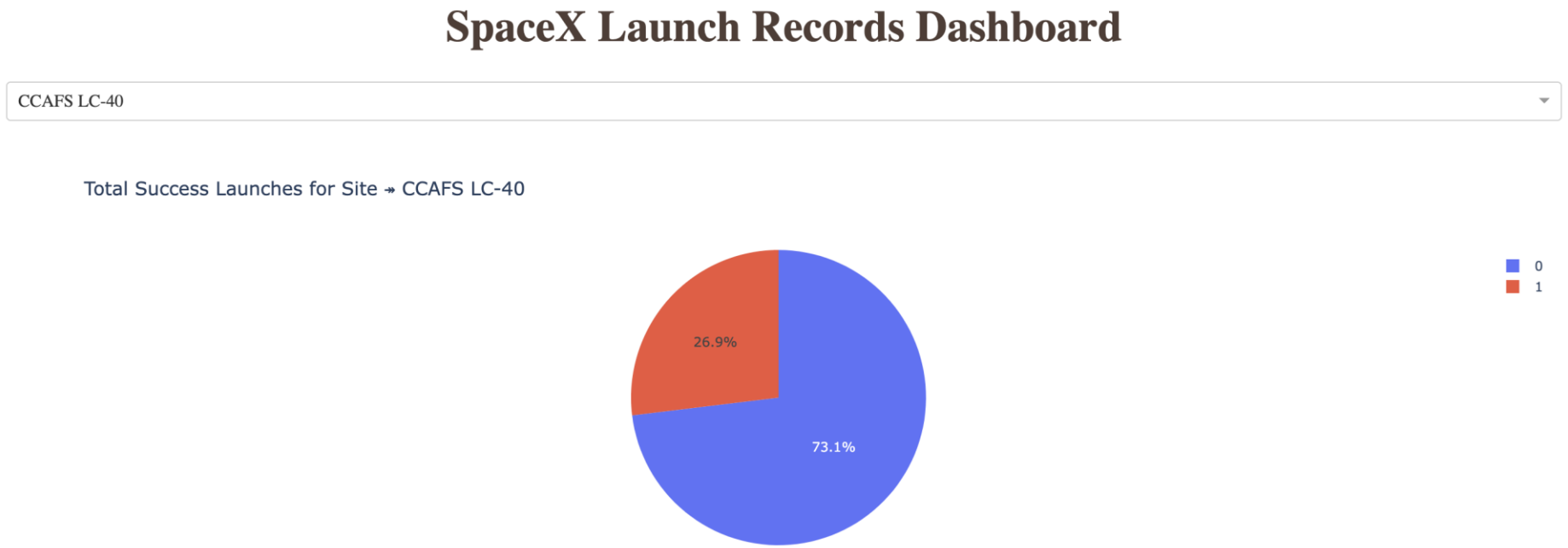
- The distances from a launch site to nearby landmarks such as the closest city, railway, or highway. The image below illustrates the distance from the VAFB SLC-4E launch site to the nearest coastline.



Data Presentation

4. Dash

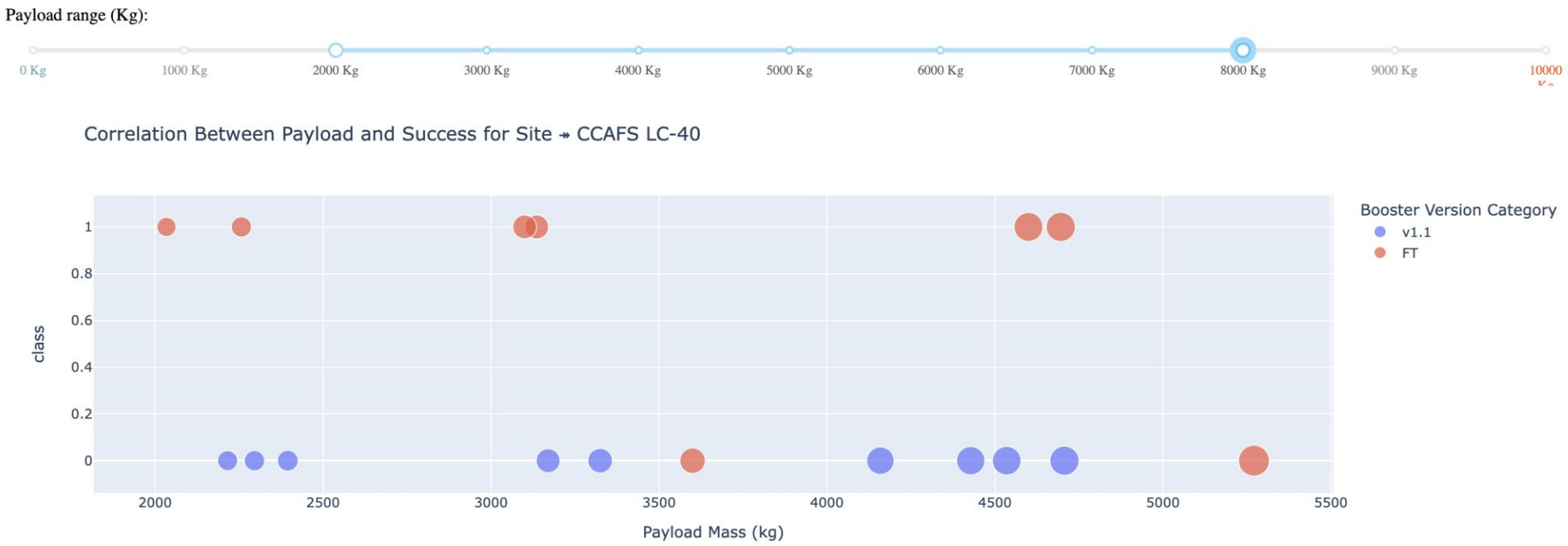
- The image below displays a pie chart for the launch site CCAFS LC-40. In this chart, 0 indicates failed launches and 1 indicates successful launches. It reveals that 73.1% of the launches at CCAFS LC-40 have been unsuccessful.



Data Presentation

4. Dash

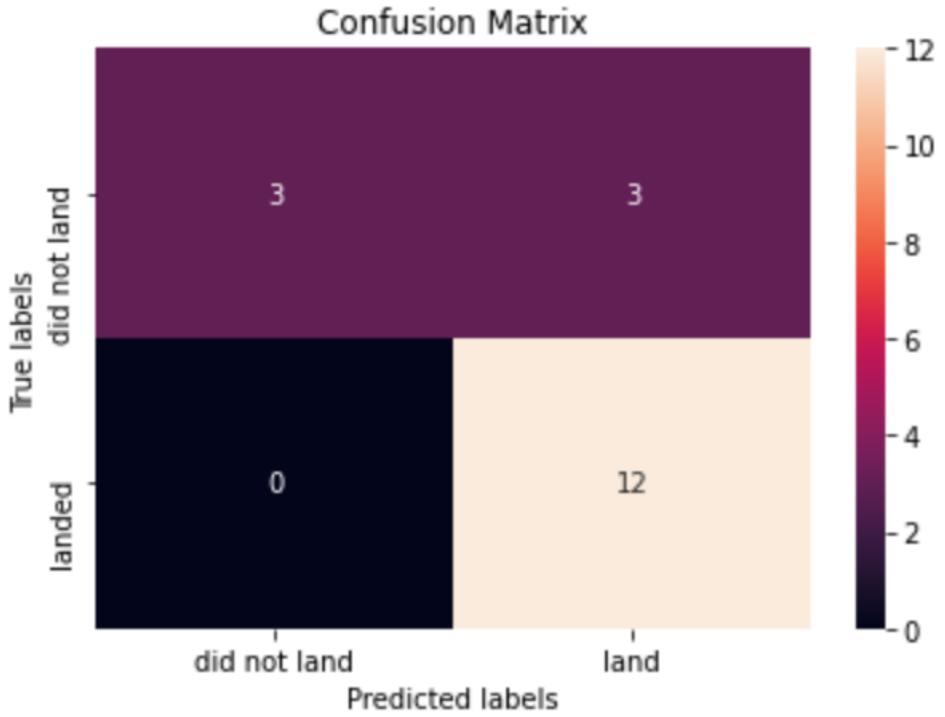
- The image below displays a scatterplot for payload masses ranging from 2000kg to 8000kg. In this scatterplot, Class 0 represents failed launches, and Class 1 represents successful launches.



Data Presentation

5. Predictive Analysis

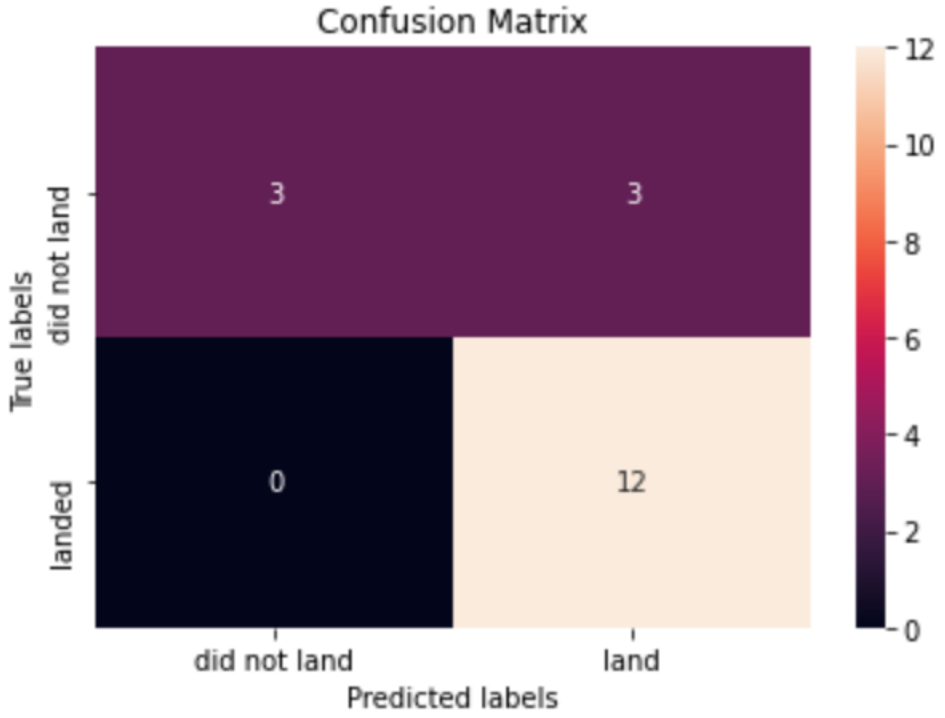
- Logistic Regression Results:
  - GridSearchCV Best Score: 0.8464285714285713
  - Accuracy Score on Test Set: 0.8333333333333334
  - Confusion Matrix:



Data Presentation

5. Predictive Analysis

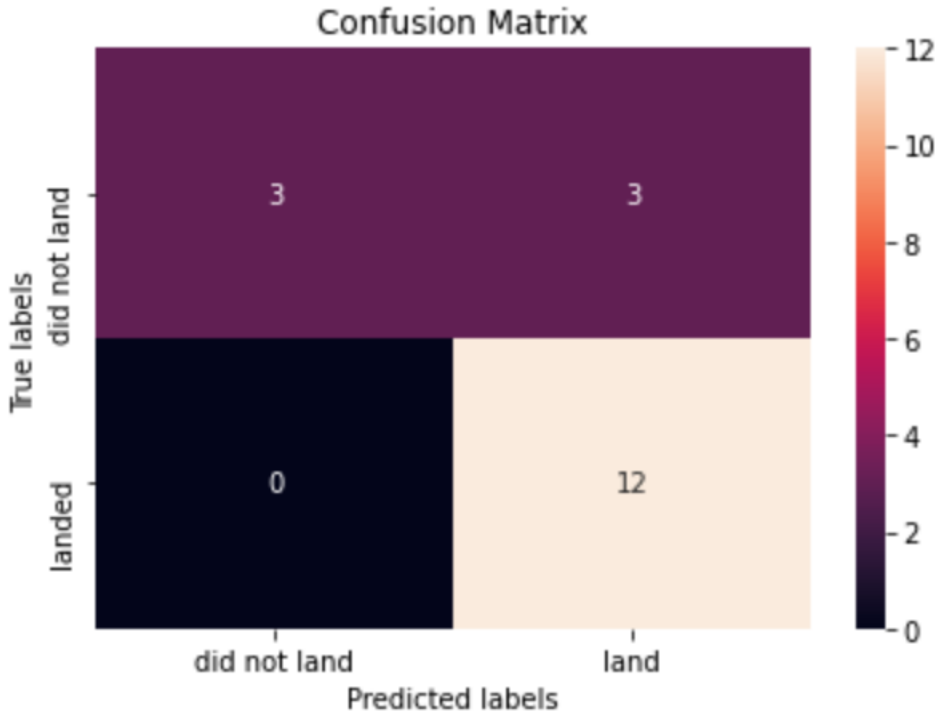
- Support Vector Machine (SVM) Results:
  - GridSearchCV Best Score: 0.8482142857142856
  - Accuracy Score on Test Set: 0.8333333333333334
  - Confusion Matrix:



Data Presentation

5. Predictive Analysis

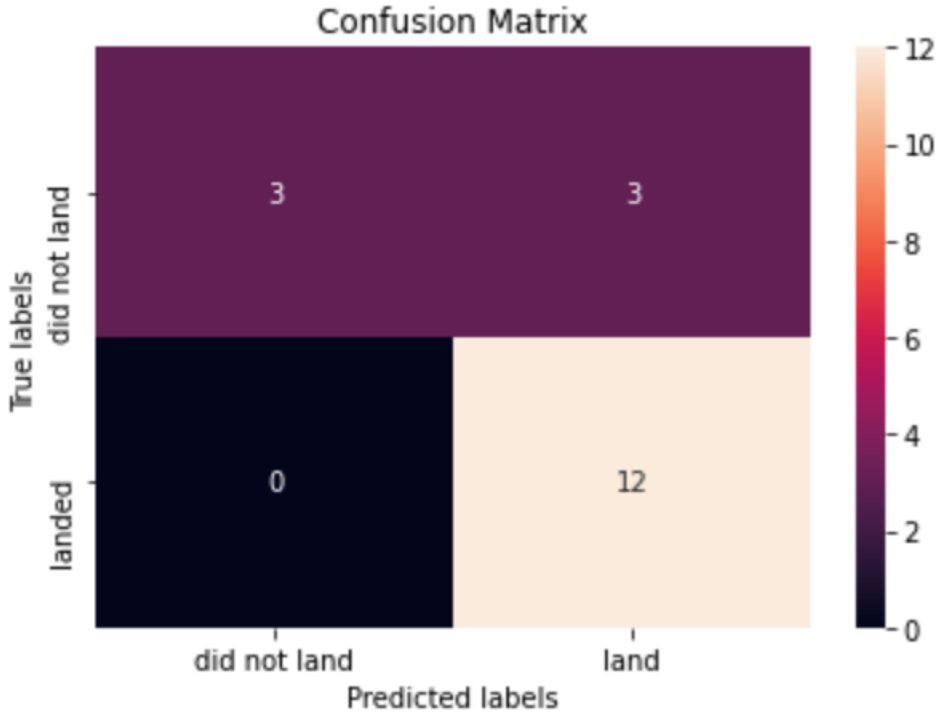
- Decision Tree Results:
  - GridSearchCV Best Score: 0.8892857142857142
  - Accuracy Score on Test Set: 0.8333333333333334
  - Confusion Matrix:



Data Presentation

5. Predictive Analysis

- K-Nearest Neighbors (KNN) Results:
  - GridSearchCV Best Score: 0.8482142857142858
  - Accuracy Score on Test Set: 0.8333333333333334
  - Confusion Matrix:



Analysis

- From the data visualization section, it is evident that certain features may correlate with mission outcomes in various ways. For instance, heavier payloads tend to have higher successful landing rates for orbit types such as Polar, LEO, and ISS. However, for GTO orbits, the distinction is less clear, as both successful and unsuccessful landing rates are observed.

- Thus, each feature may influence the final mission outcome to some extent. While the specific impact of each feature on the mission outcome can be complex and not immediately apparent, machine learning algorithms can be employed to discern patterns in historical data. These algorithms can then predict the likelihood of a mission's success based on the given features.

Wrap-Up

- In this project, we aim to predict whether the first stage of a given Falcon 9 launch will successfully land, which in turn helps determine the cost of the launch. Each feature of a Falcon 9 launch, such as payload mass or orbit type, can influence the mission outcome. We employed several machine learning algorithms to analyze past Falcon 9 launch data and develop predictive models.

- Among the four machine learning algorithms tested, the predictive model produced by the decision tree algorithm demonstrated the highest performance.