Applied Data Science capstone

Todd Nicholas

June 10th, 2024

Background image: https://www.google.com/url?sa=i&url=https%3A%2F%2Fspacenews.com%2Fnext-commercial-falcon-heavy-mission-to-launch-debut-astranis-satellite%2F&psig=AOvVaw2Fk66iSj6PnCc2akAiFPco&ust=1641344747041000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCLi42-bzlvUCFQAAAAAdAAAAABAG

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion

EXECUTIVE SUMMARY

In this capstone project, we aim to predict the success of the SpaceX Falcon 9 first stage landings using various machine learning classification algorithms. The project involves several key steps:

1. Data Collection, Wrangling, and Formatting: We gather and preprocess the data to ensure it is clean and suitable for analysis.
2. Exploratory Data Analysis: We perform a thorough examination of the data to identify patterns and correlations.
3. Interactive Data Visualization: We create visual tools to interactively explore the data and its relationships.
4. Machine Learning Prediction: We apply multiple machine learning algorithms to predict the landing outcome of the Falcon 9 first stage.

Our analysis shows that certain features of the rocket launches are correlated with their outcomes, indicating factors that influence success or failure.

INTRODUCTION

In this capstone project, we aim to predict the success of the Falcon 9 first stage landings. SpaceX advertises its Falcon 9 rocket launches at $62 million, significantly lower than other providers, which can cost upwards of $165 million. A key factor in these savings is SpaceX's ability to reuse the first stage. Thus, predicting the success of the first stage landing can help determine the launch cost, providing crucial information for companies considering bidding against SpaceX for rocket.

The project includes data collection, wrangling, and formatting to prepare the data for analysis. We will conduct exploratory data analysis to identify patterns and correlations, create interactive visualizations, and apply various machine learning algorithms to predict landing success. Our preliminary findings suggest that the decision tree algorithm may be the most effective model for this prediction task.

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)

METHODOLOGY

1 Data collection, wrangling, and formatting

The SpaceX API, accessed via https://api.spacexdata.com/v4/rockets/, delivers detailed information on various rocket launches conducted by SpaceX. For our specific analysis, we filtered the data to focus exclusively on Falcon 9 launches.

To address any missing values in the dataset, we replaced them with the mean of their respective columns. This process resulted in a dataset comprising 90 rows (instances) and 17 columns (features). The image below illustrates the initial few rows of this curated dataset.

	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

METHODOLOGY

1 Data collection, wrangling, and formatting

The data was obtained through web scraping, specifically from the Wikipedia page "List of Falcon 9 and Falcon Heavy launches" using the URL: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=.
This webpage contains detailed information exclusively about Falcon 9 launches. After processing, the dataset comprises 121 rows (instances) and 11 columns (features). The image below depicts the first few rows of this dataset.

	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

METHODOLOGY

1 Data collection, wrangling, and formatting

The data undergoes further processing to eliminate any missing entries, and categorical features are transformed using one-hot encoding. An additional column named ‘Class’ is introduced to the dataframe, indicating the outcome of each launch: 0 for a failed launch and 1 for a successful one. After these adjustments, the refined dataset consists of 90 rows (instances) and 83 columns (features).

METHODOLOGY

2 Exploratory Data Analysis (EDA)

Pandas and NumPy

We utilize functions from the Pandas and NumPy libraries to extract and analyze essential information from the collected data. This analysis includes:

- Determining the number of launches at each launch site.
- Counting the occurrences of each orbit type.
- Enumerating the outcomes of each mission.

SQL

SQL queries are employed to delve deeper into the data and answer specific questions, such as:

- Identifying the unique launch site names involved in the space missions.
- Calculating the total payload mass carried by boosters launched by NASA (CRS).
- Finding the average payload mass carried by the booster version F9 v1.1.

METHODOLOGY

3 Data Visualization

Matplotlib and Seaborn

We employ functions from the Matplotlib and Seaborn libraries to create various visualizations, such as scatterplots, bar charts, and line charts, to gain insights into the data. These visualizations help us explore relationships between different features, including:

- The correlation between flight numbers and launch sites.
- The association between payload mass and launch sites.
- The success rate in relation to different orbit types.

Folium

The Folium library is utilized to create interactive maps, allowing for a spatial understanding of the data. This includes:

Marking all launch sites on a map.Indicating successful and failed launches for each site. Highlighting distances between a launch site and nearby features such as cities, railways, or highways.

METHODOLOGY

3 Data Visualization

- Dash

We use functions from Dash to create an interactive platform that allows users to adjust inputs via a dropdown menu and a range slider. This dynamic site features visualizations such as pie charts and scatterplots to display:

- The total number of successful launches from each launch site.
- The relationship between payload mass and mission outcomes (success or failure) for each launch site.

METHODOLOGY

4 Machine Learning Prediction

We utilize functions from the Scikit-learn library to develop our machine learning models. The prediction phase involves several steps: standardizing the data, splitting it into training and test sets, and creating various machine learning models, including logistic regression, support vector machines (SVM), decision trees, and k-nearest neighbors (KNN). We then fit these models to the training set and optimize hyperparameters to find the best combination for each model. Finally, we evaluate the models based on their accuracy scores and confusion matrices. This comprehensive approach ensures robust model development and thorough evaluation, leveraging the extensive tools provided by Scikit-learn.

RESULTS

- The results are split into 5 sections:
  - SQL (EDA with SQL)
  - Matplotlib and Seaborn (EDA with Visualization)
  - Folium
  - Dash
  - Predictive Analysis
- In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

RESULTS

1 SQL (EDA with SQL)

- The names of the unique launch sites in the space mission

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

- 5 records where launch sites begin 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

RESULTS

1 SQL (EDA with SQL)

- The total payload mass carried by boosters launched by NASA (CRS)

Total payload mass by NASA (CRS)
45596

- The average payload mass carried by booster version F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

- The date when the first successful landing outcome in ground pad was

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

RESULTS

1 SQL (EDA with SQL)

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

- The total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

RESULTS

1 SQL (EDA with SQL)

- The names of the booster versions which have carried the maximum payload mass

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

RESULTS

1 SQL (EDA with SQL)

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

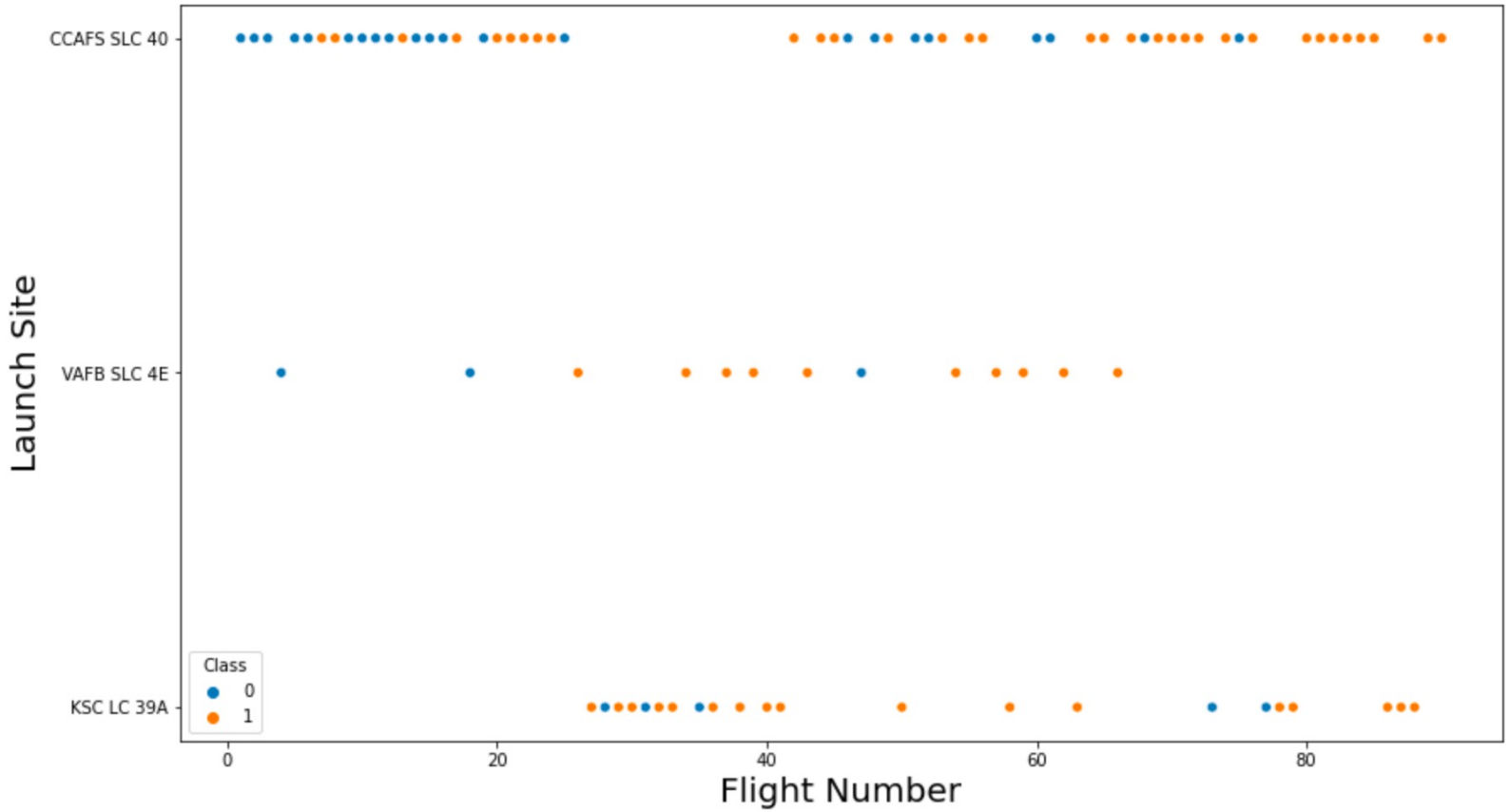
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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

RESULTS

② Matplotlib and Seaborn (EDA with Visualization)

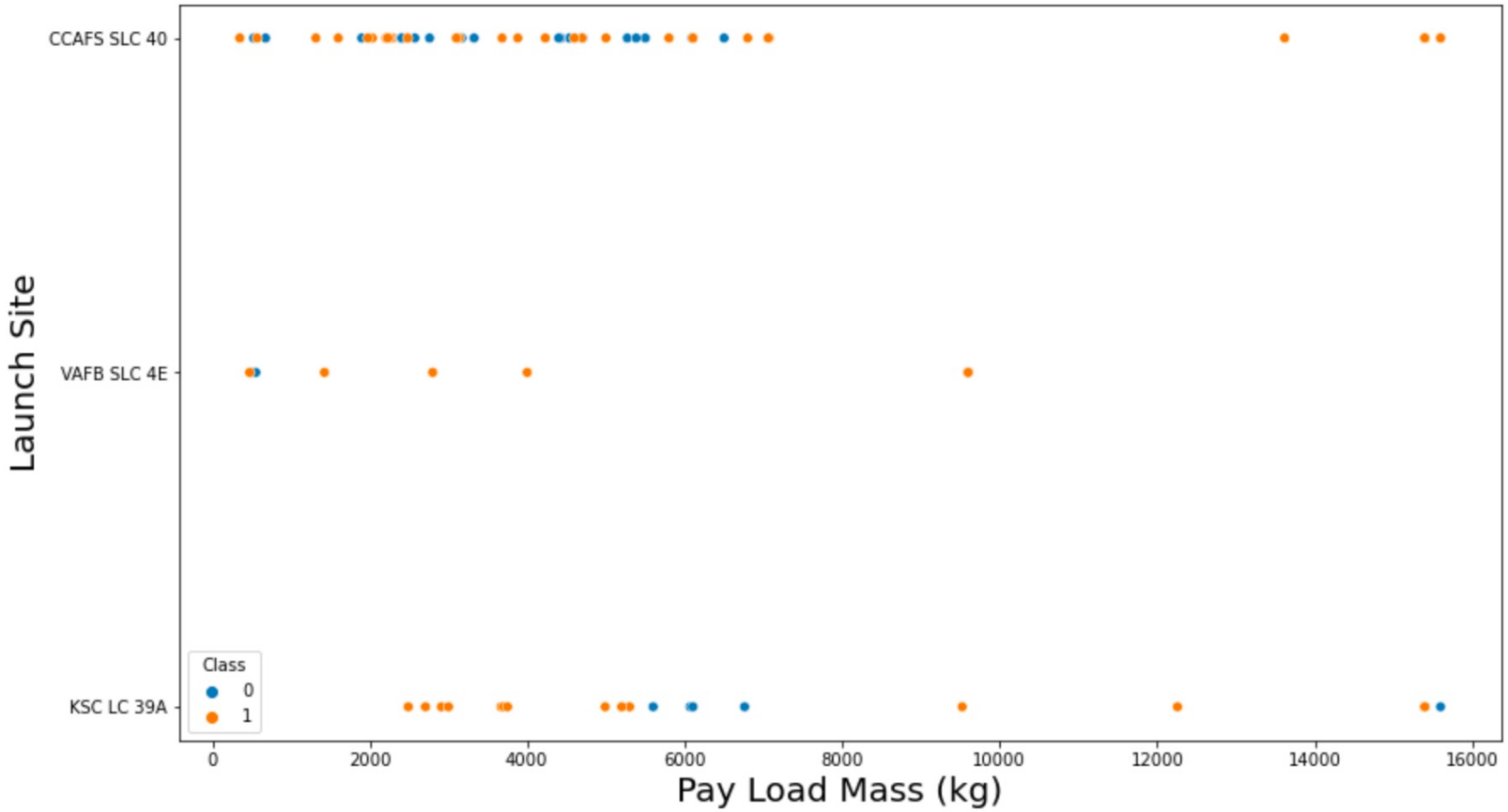
- The relationship between flight number and launch site



RESULTS

② Matplotlib and Seaborn (EDA with Visualization)

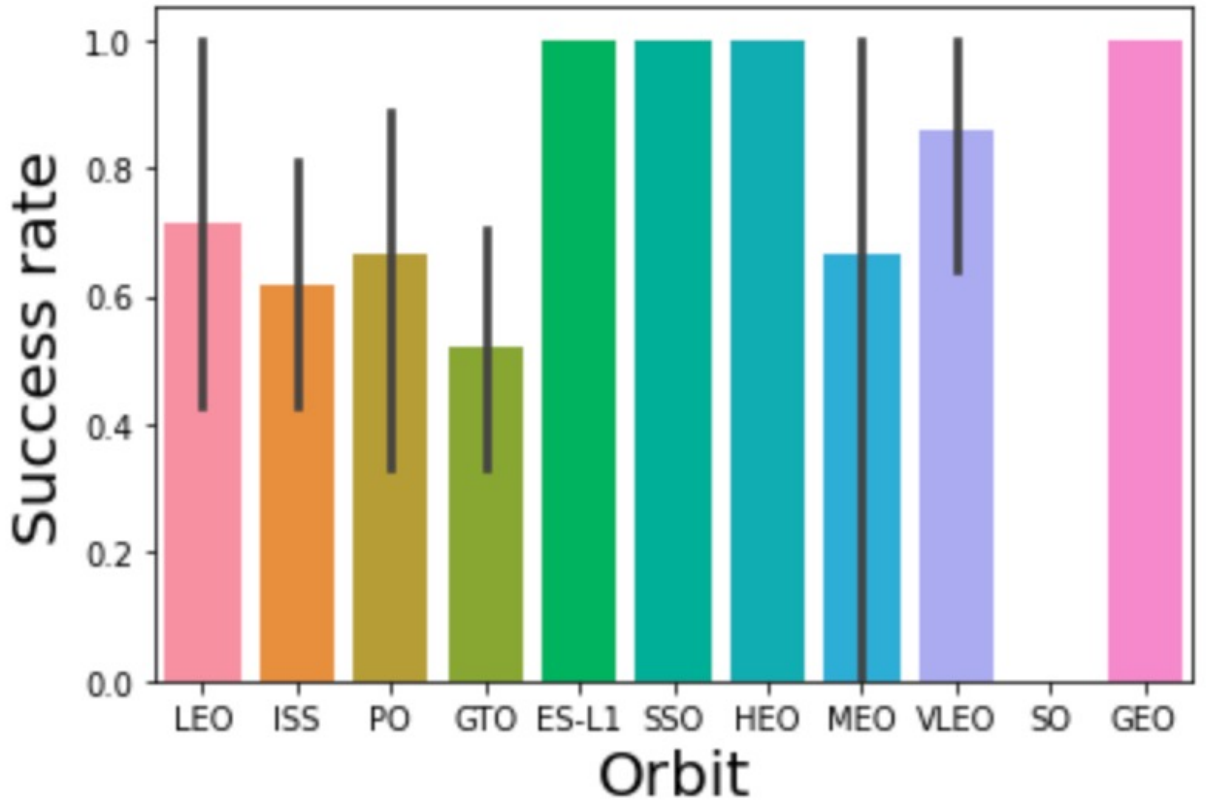
- The relationship between payload mass and launch site



RESULTS

② Matplotlib and Seaborn (EDA with Visualization)

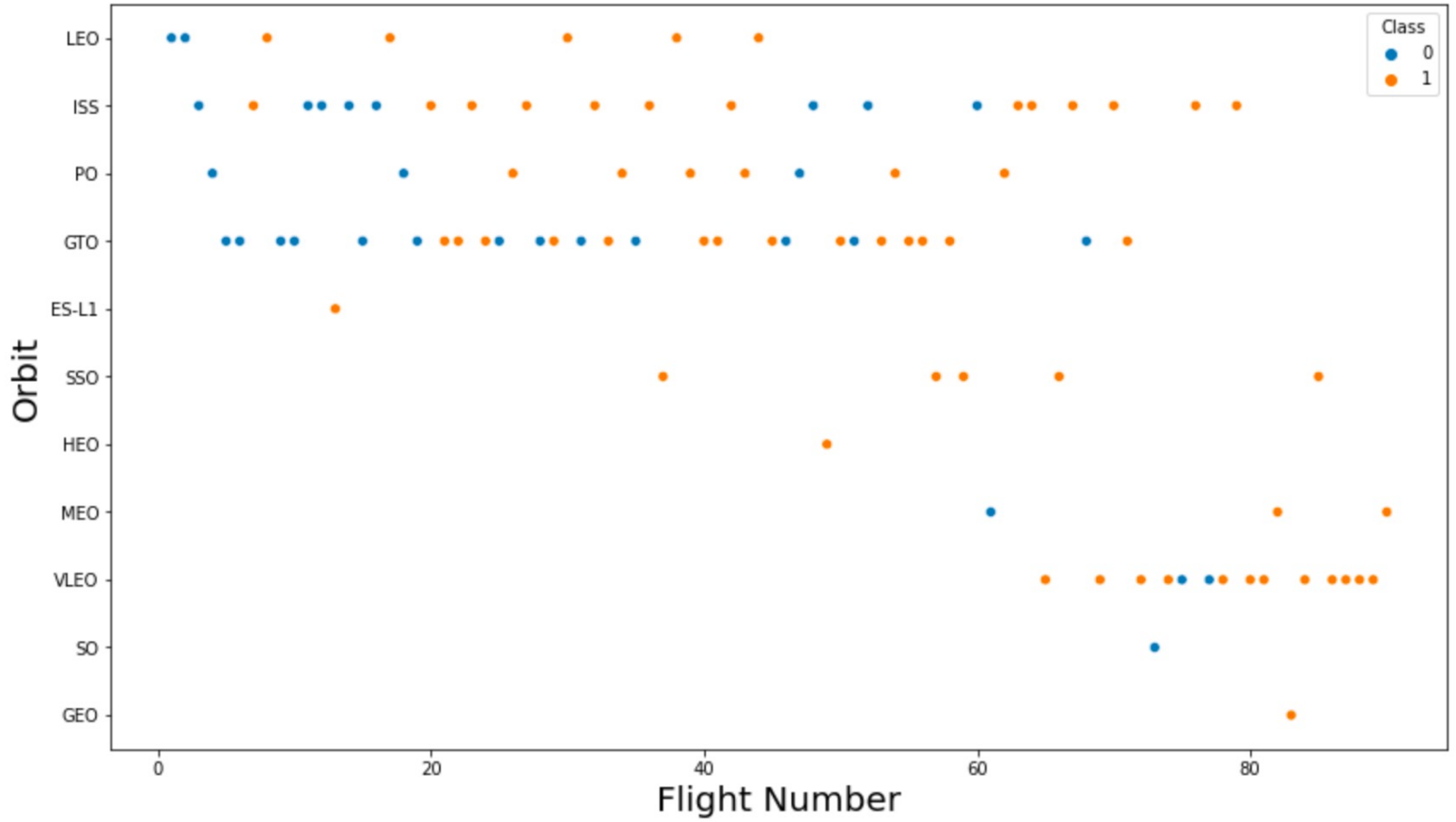
- The relationship between success rate and orbit type



RESULTS

② Matplotlib and Seaborn (EDA with Visualization)

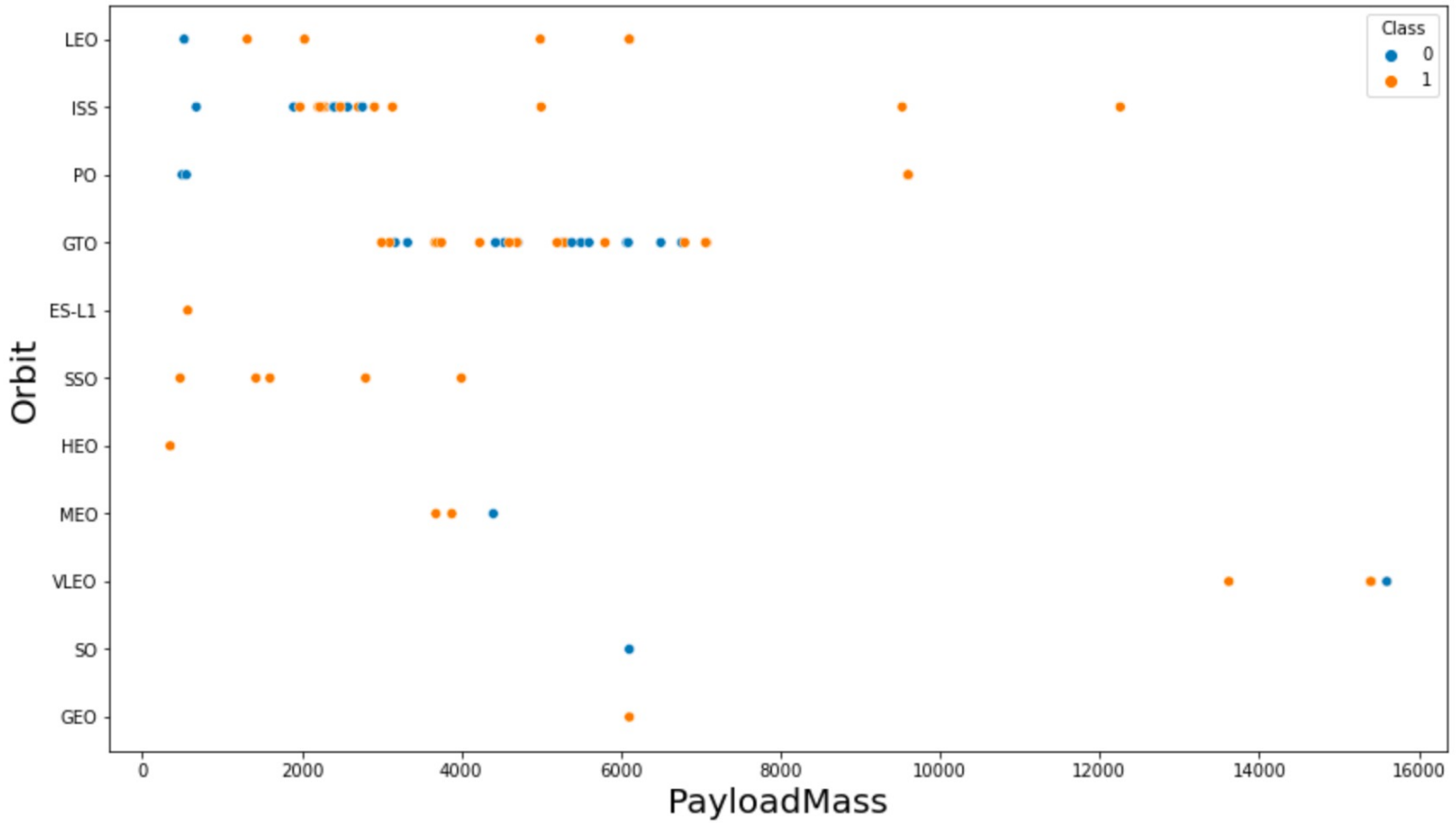
- The relationship between flight number and orbit type



RESULTS

② Matplotlib and Seaborn (EDA with Visualization)

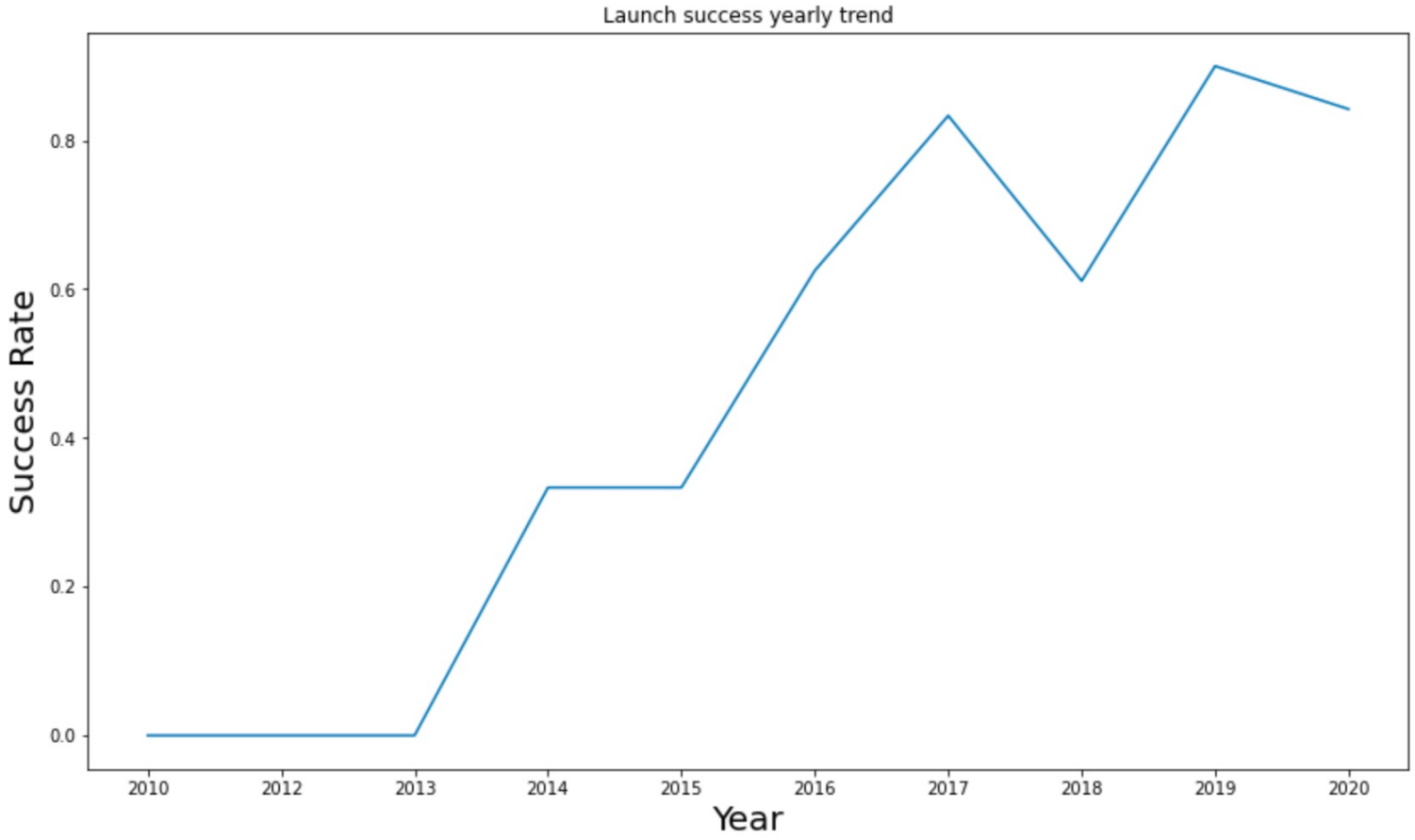
- The relationship between payload mass and orbit type



RESULTS

② Matplotlib and Seaborn (EDA with Visualization)

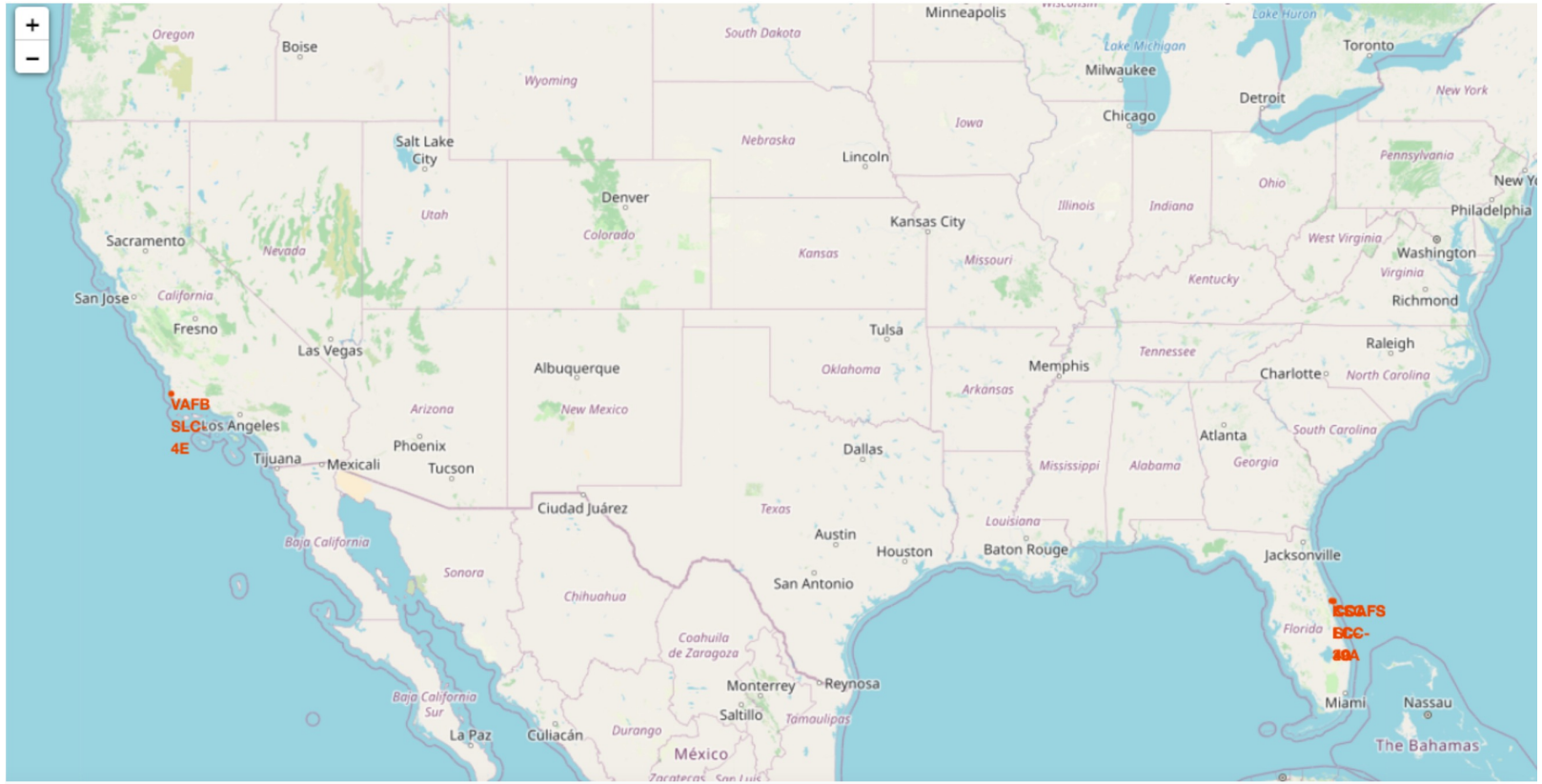
- The launch success yearly trend



RESULTS

3 Folium

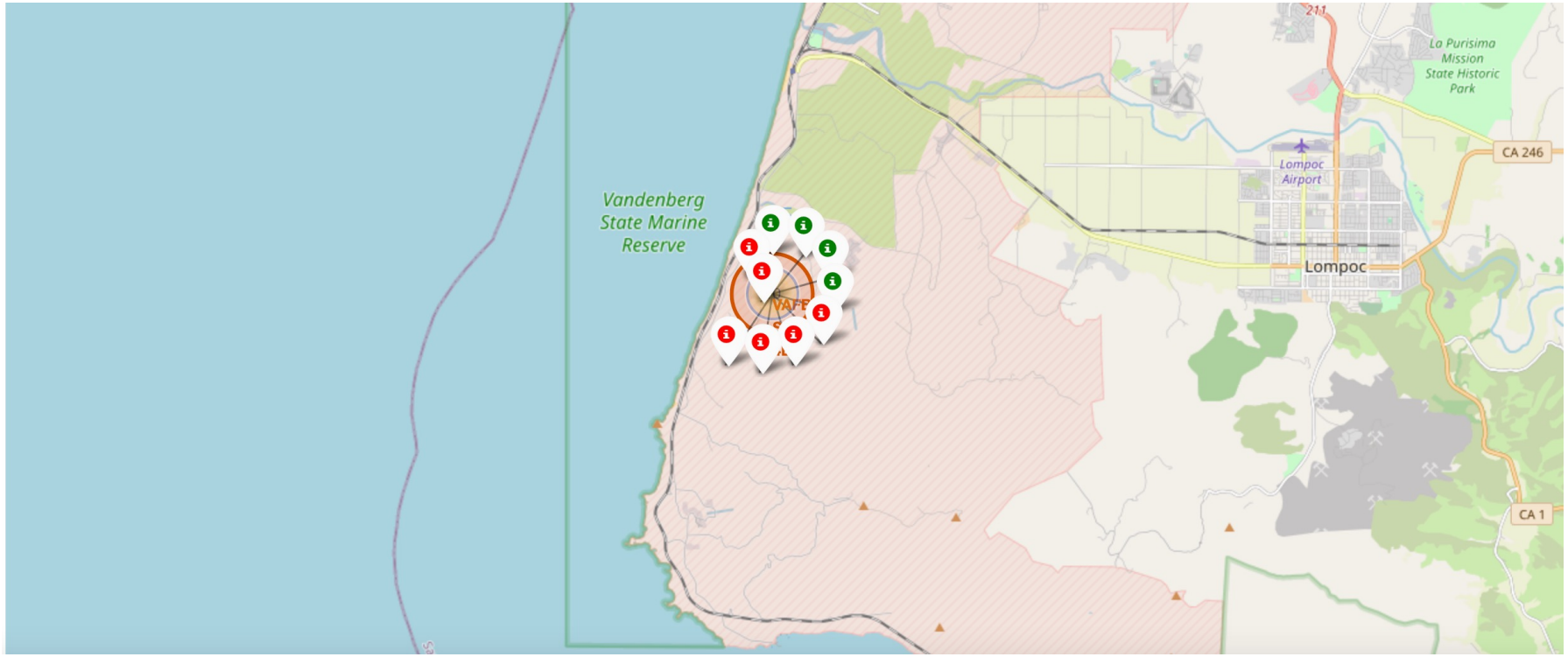
- All launch sites on map



RESULTS

3 Folium

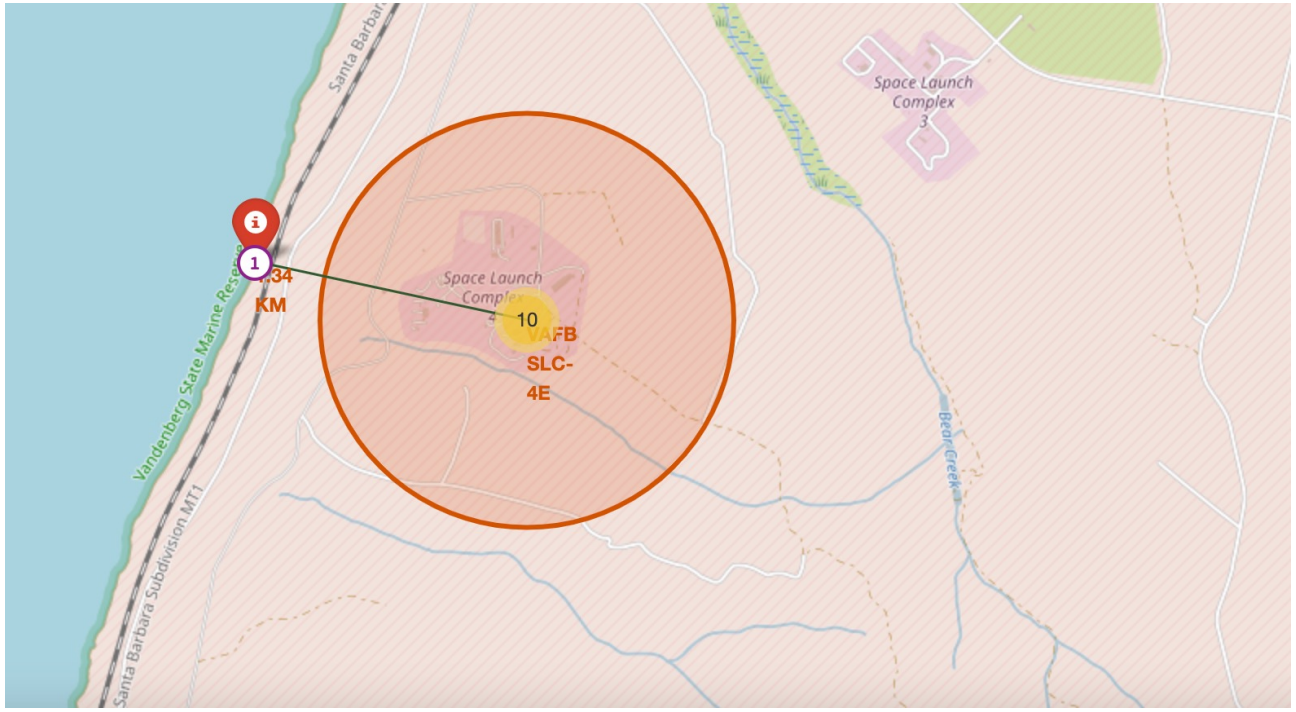
- The succeeded launches and failed launches for each site on map
  - If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch



RESULTS

3 Folium

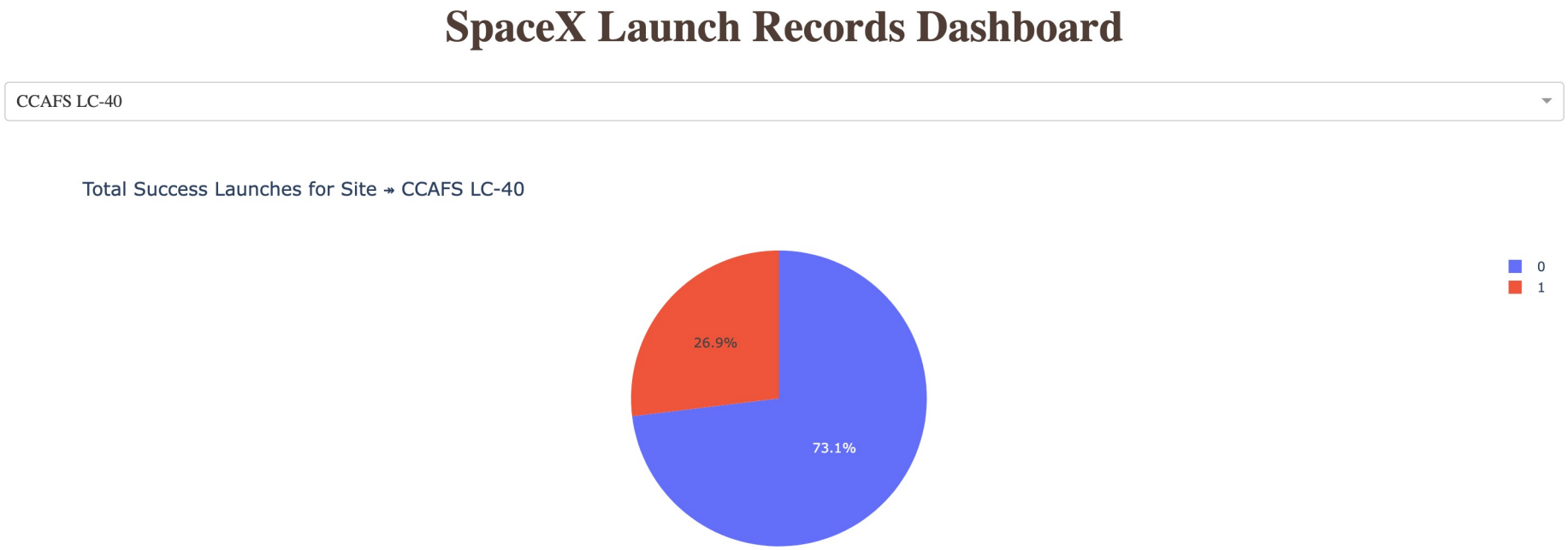
- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



RESULTS

4 Dash

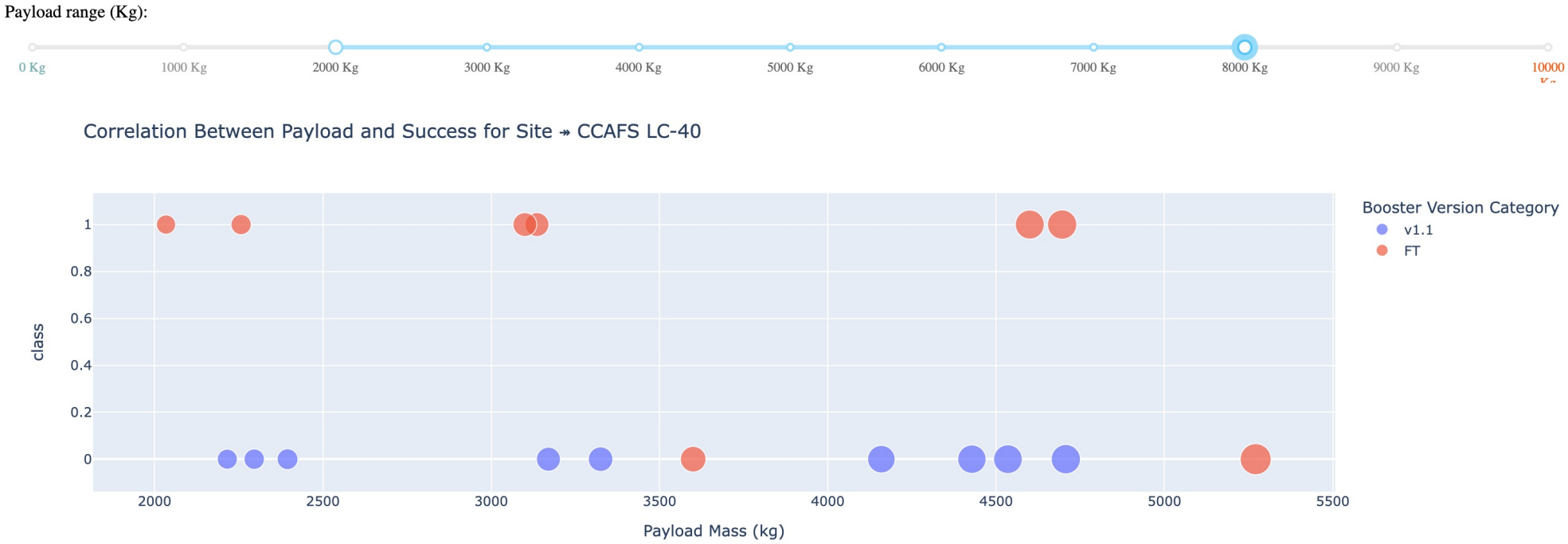
- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.



RESULTS

4 Dash

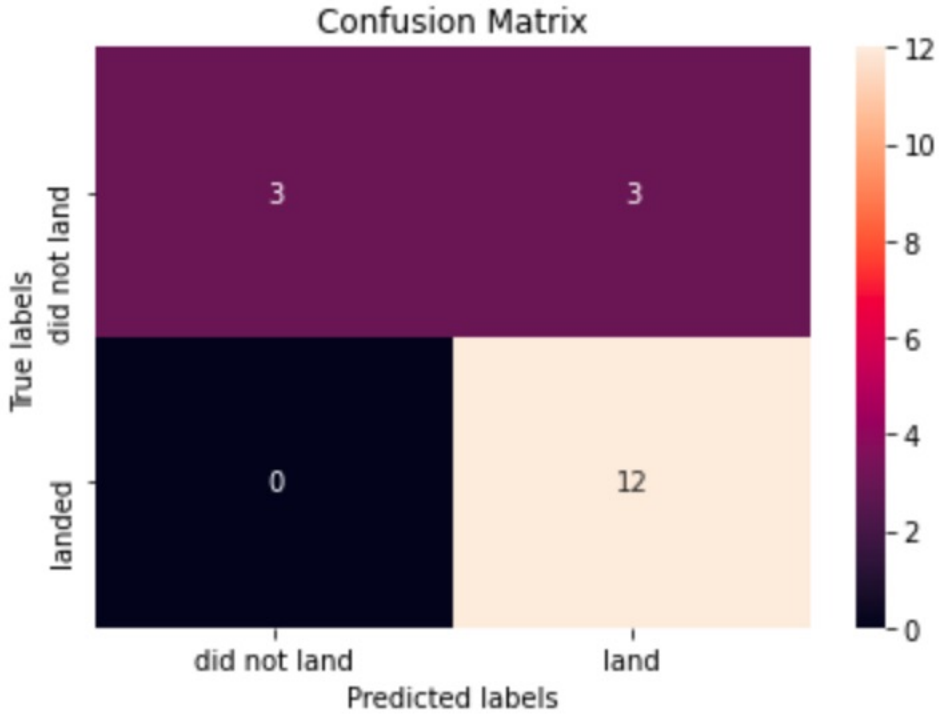
- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.



RESULTS

5 Predictive Analysis

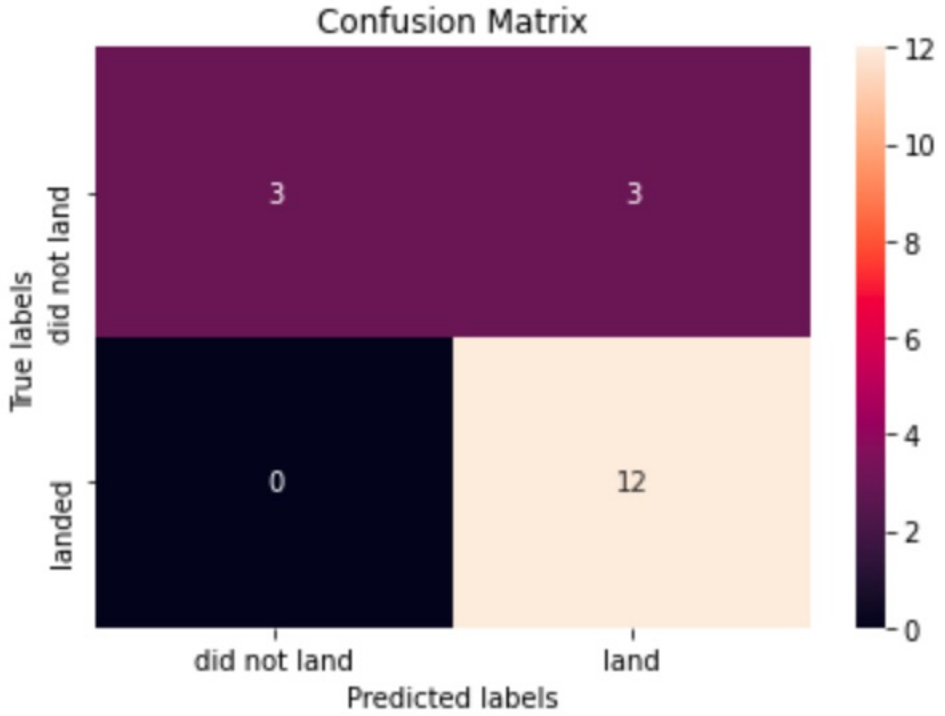
- Logistic regression
  - GridSearchCV best score: 0.8464285714285713
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

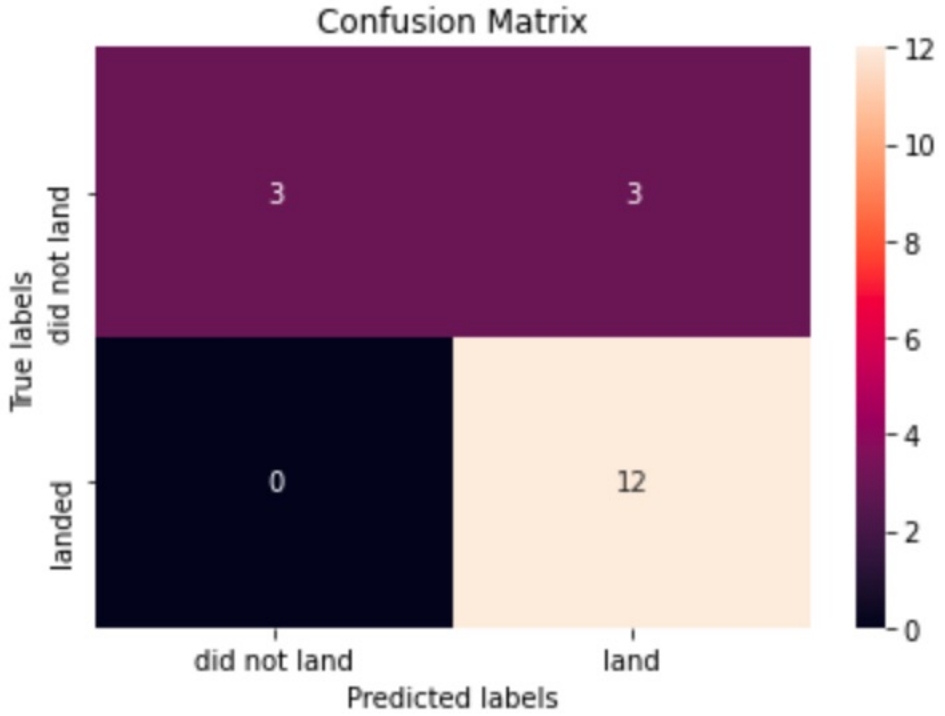
- Support vector machine (SVM)
  - GridSearchCV best score: 0.8482142857142856
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

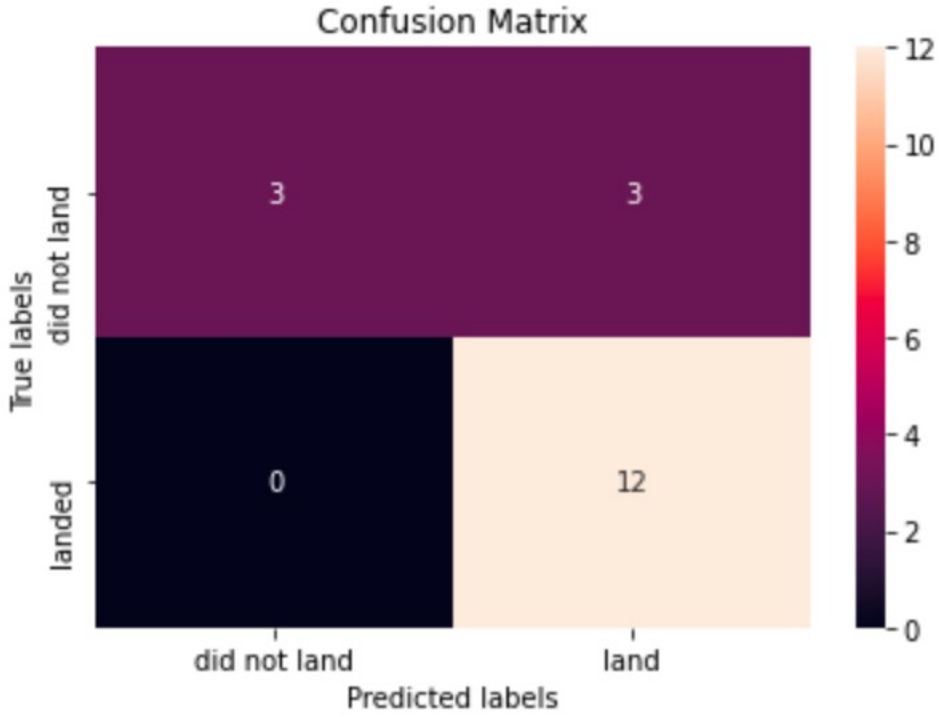
- Decision tree
  - GridSearchCV best score: 0.8892857142857142
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

- K nearest neighbors (KNN)
  - GridSearchCV best score: 0.8482142857142858
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

Comparing the results of all four models side by side reveals that they share the same accuracy score and confusion matrix on the test set. Therefore, we rank them based on their GridSearchCV best scores. The models are ranked as follows, from best to worst:

1. Decision Tree: GridSearchCV best score of 0.8893
2. K-Nearest Neighbors (KNN): GridSearchCV best score of 0.8482
3. Support Vector Machine (SVM): GridSearchCV best score of 0.8482
4. Logistic Regression: GridSearchCV best score of 0.8464

DISCUSSION

From the data visualization section, we observe that certain features may correlate with the mission outcome in various ways. For instance, heavy payloads tend to have higher successful landing rates for orbit types such as Polar, LEO, and ISS. However, for GTO, the distinction is less clear as both positive and negative landing rates are present.

Thus, each feature likely impacts the final mission outcome, though the specific influences of these features are challenging to pinpoint. To uncover these patterns and predict mission success, we can employ machine learning algorithms to analyze past data and identify the relationships between features and mission outcomes.

CONCLUSION

In this project, we aim to predict whether the first stage of a given Falcon 9 launch will successfully land, which is crucial for determining the overall launch cost. Each feature of a Falcon 9 launch, such as payload mass or orbit type, may influence the mission outcome.

We employed several machine learning algorithms to analyze past Falcon 9 launch data and develop predictive models. Among the four algorithms used, the decision tree model achieved the best performance.