Applied Data Science Capstone

EXECUTIVE SUMMARY

The capstone project aims to predict the landing outcome of SpaceX's Falcon 9 first stage with various machine learning classifiers.

The project's core activities are:

- Gathering, refining, and structuring data
- Initial data examination
- Creating dynamic visualizations of data
- Prediction through machine learning

Our analysis indicates a link between certain launch characteristics and the final result, whether a launch succeeds or fails. We've also deduced that a decision tree could be the optimal predictive model for determining the success of the Falcon 9's first stage landing.

INTRODUCTION

In this capstone project, our objective is to forecast the successful landing of SpaceX's Falcon 9 first stage. While SpaceX lists a launch price of $62 million, competitors may charge over $165 million. The ability to reuse the first stage underpins SpaceX's cost advantage. By predicting landing outcomes, we can estimate launch costs, valuable intelligence for potential SpaceX competitors in bidding processes.

Although some Falcon 9 landings fail as part of planned disposals into the ocean, our analysis focuses on the correlation between launch attributes—such as payload mass, orbit type, and launch site—and the likelihood of a successful first-stage landing.

METHODOLOGY

Our capstone project methodology is structured as follows:

1. For data collection, cleaning, and preparation, we utilize the SpaceX API and methods of web scraping.

2. We conduct Exploratory Data Analysis (EDA) with the aid of libraries such as Pandas, NumPy, and tools like SQL for database interaction.

3. We employ Matplotlib, Seaborn, Folium, and Dash for the visualization of data to understand and present our findings visually.

4. Lastly, we apply machine learning algorithms for prediction purposes, including Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).

METHODOLOGY

1. Gathering, refining, and structuring data

In the first phase of our methodology, we concentrate on sourcing and refining data. We utilize the SpaceX API, available at https://api.spacexdata.com/v4/rockets/, to collect information on a range of SpaceX rocket launches, specifically narrowing down to Falcon 9 launches.

To ensure data quality, we address missing values by imputing the mean of the respective feature they are missing from. The dataset we compile consists of 90 instances, each described by 17 features. An example of the initial data can be seen in the subsequent illustration.

	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. Gathering, refining, and structuring data

Additionally, we employ web scraping to augment our dataset with Falcon 9 launch details from Wikipedia, specifically from the page "List of Falcon 9 and Falcon Heavy launches," archived at https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922.

This source exclusively provides Falcon 9 launch data, from which we extracted a dataset comprising 121 records, each characterized by 11 attributes. A depiction of the initial dataset rows is provided in the following visual.

	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. Gathering, refining, and structuring data

Subsequent to the initial data collection, the dataset undergoes further processing to fill any missing values and encode categorical features using one-hot encoding technique. We also introduce an additional 'Class' column to the dataset, which assigns a '0' for a failed launch and a '1' for a successful one. This results in a refined dataset containing 90 instances and 83 features.

METHODOLOGY

2. Exploratory Data Analysis (EDA)

Within our exploratory data analysis, we leverage Pandas and NumPy to extract foundational insights from our data, including the frequency of launches per site, the distribution of orbits, and the success rates of missions.

For more in-depth queries, we use SQL to address questions regarding the dataset, such as identifying all unique launch sites, computing the total and average payload mass for specific missions like those under NASA's CRS program, and assessing payload variations across different versions of the Falcon 9 rocket, like the F9 v1.1.

METHODOLOGY

2. Exploratory Data Analysis (EDA)

For data visualization, we utilize Matplotlib and Seaborn to create scatterplots, bar charts, and line charts, which help us interpret the relationships between key features such as flight number versus launch site, payload mass versus launch site, and success rate versus orbit type.

Additionally, we employ Folium for interactive mapping to pinpoint all launch sites, differentiate between successful and unsuccessful launches at each site, and display the proximity of launch sites to nearby cities, railways, or highways.

METHODOLOGY

3. Creating dynamic visualizations of data

We use Dash to construct an interactive web application that enables users to manipulate the input through dropdown menus and a range slider. The application features a pie chart and a scatterplot that reveal the aggregate number of successful launches by launch site and explore the relationship between payload mass and mission outcomes, distinguishing between success and failure for each launch site.

Dash

by plotly

METHODOLOGY

4. Prediction through machine learning

We utilize Scikit-learn's suite of tools to construct our machine learning models. Our predictive modeling process encompasses data normalization, partitioning the dataset into training and testing subsets, and formulating various machine learning algorithms such as Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).

We train these models, optimize their hyperparameters, and assess their performance using accuracy metrics and confusion matrices.

RESULTS

Our findings are organized into five distinct sections:

- SQL (EDA with SQL)
- Matplotlib and Seaborn (EDA with Visualization)
- Folium
- Dash
- Predictive Analysis

For clarity in the visual representations throughout these sections, 'class 0' denotes a failed launch, whereas 'class 1' signifies a launch that was successful.

RESULTS

SQL (EDA with SQL)

The distinct launch sites utilized for the space missions are listed,

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

and we have identified five instances where the launch sites' designations 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

RESULTS

SQL (EDA with SQL)

Booster missions conducted by NASA under the CRS program have a cumulative payload mass of 45,596 kilograms.

Total payload mass by NASA (CRS)
45596

The average mass of payloads carried by the Falcon 9 version 1.1 booster is 2,928 kilograms.

Average payload mass by Booster Version F9 v1.1
2928

The first successful ground pad landing occurred on December 22, 2015.

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

RESULTS

SQL (EDA with SQL)

Boosters that have successfully landed on a drone ship and carried a payload mass between 4000 and 6000 kilograms are listed.

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

The total count of mission outcomes includes 100 successes and 1 failure.

number_of_success_outcomes	number_of_failure_outcomes
100	1

RESULTS

SQL (EDA with SQL)

Listed are the versions of Falcon 9 boosters 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

RESULTS

SQL (EDA with SQL)

In 2015, there were failed drone ship landing attempts associated with specific Falcon 9 booster versions and launch sites.

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

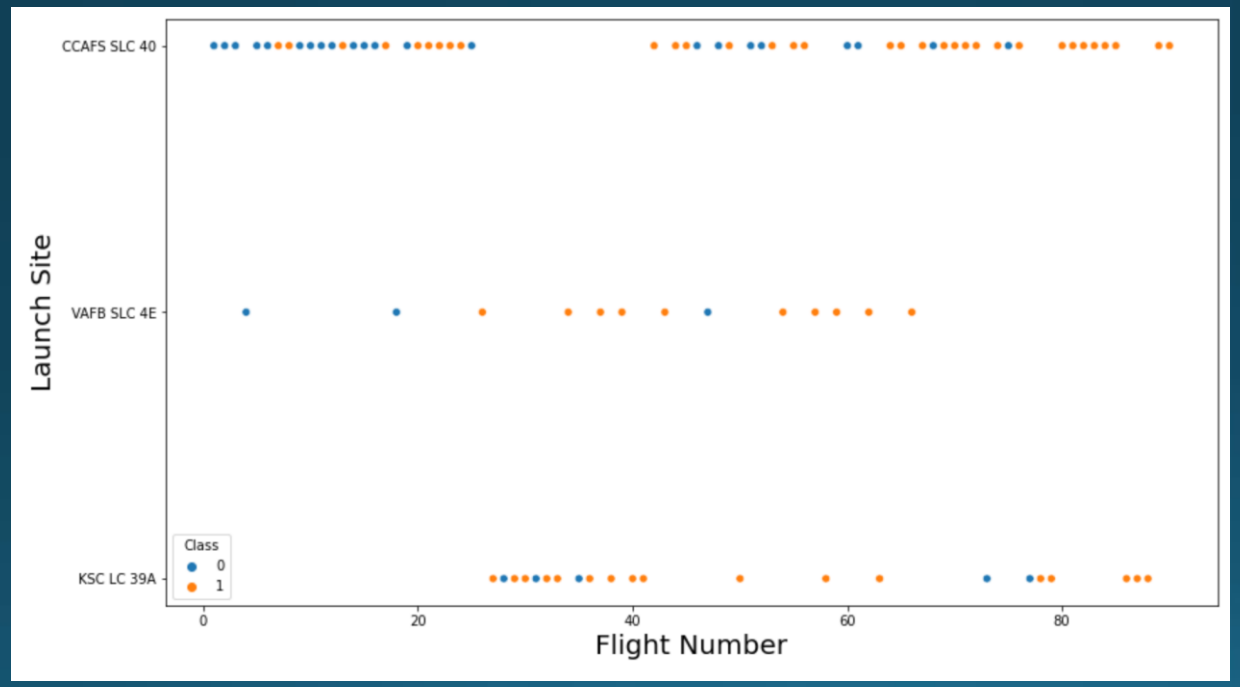
Additionally, a ranking of landing outcomes based on their frequency from 2010-06-04 to 2017-03-20 has been compiled, ordered from the most to the least occurrences.

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

Matplotlib and Seaborn (EDA with Visualization)

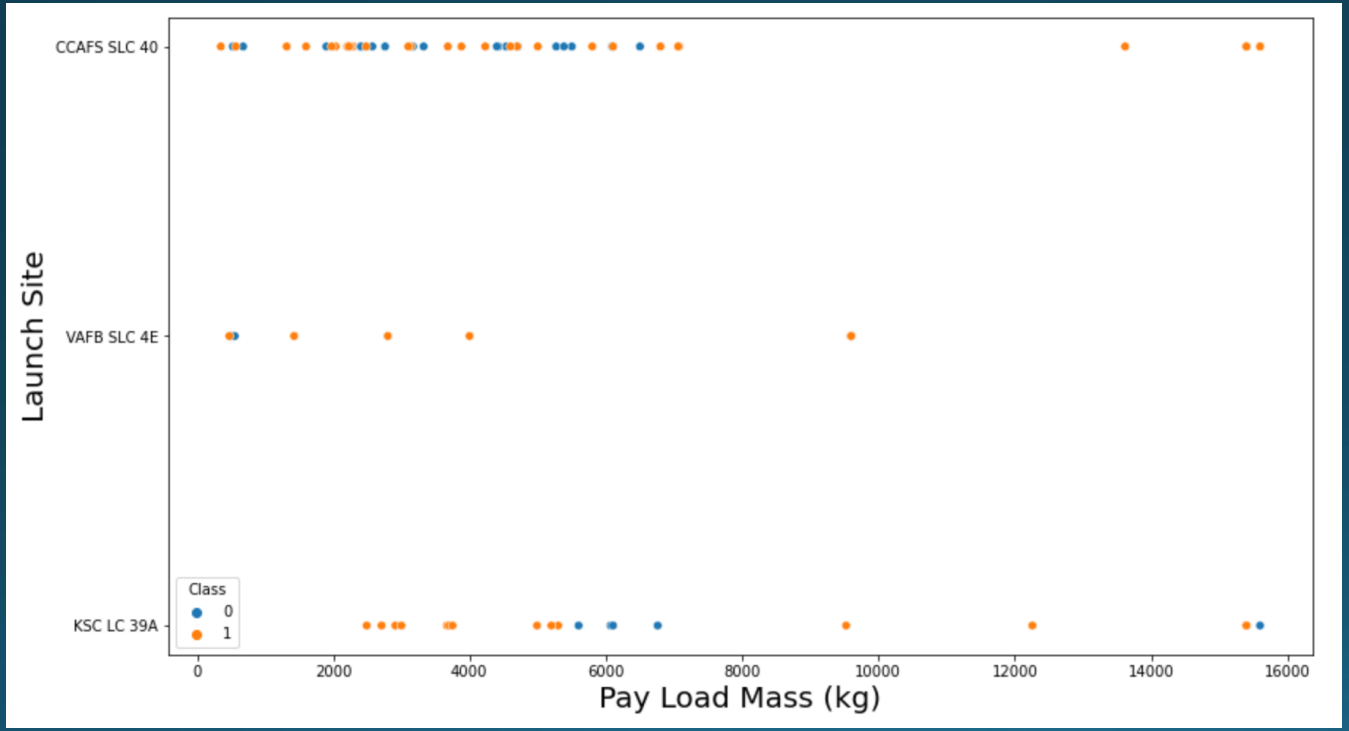
The correlation between launch numbers and their respective launch sites.



RESULTS

Matplotlib and Seaborn (EDA with Visualization)

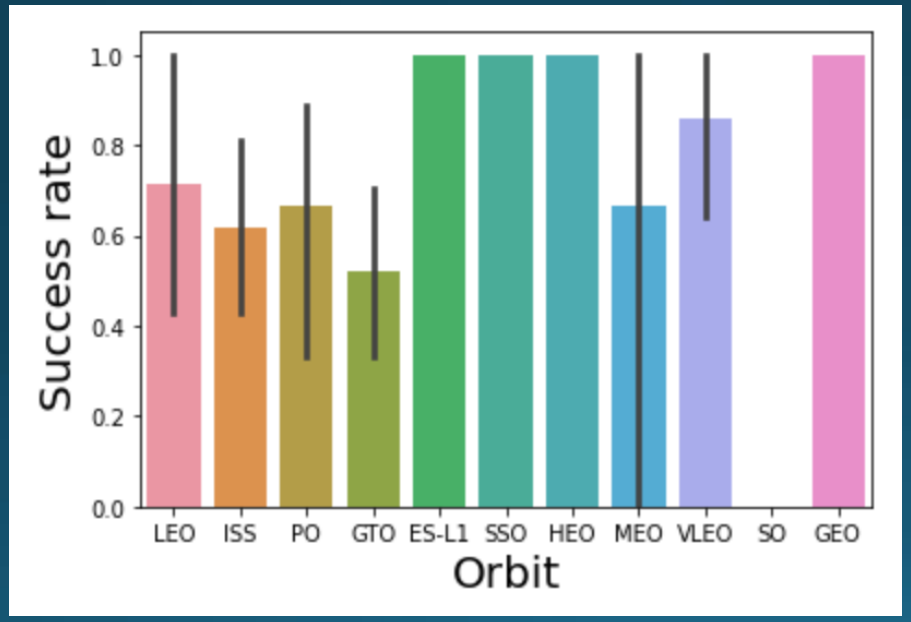
The connection between the mass of payloads and the locations from which they are launched.



RESULTS

Matplotlib and Seaborn (EDA with Visualization)

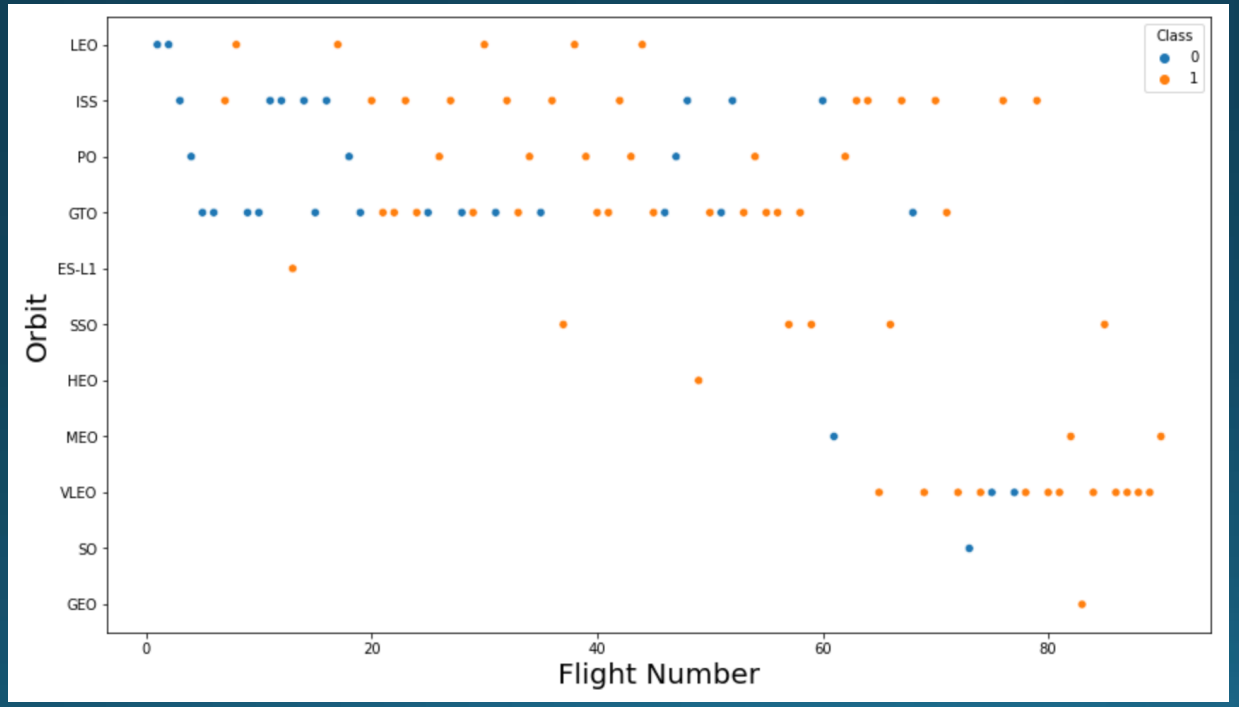
The link between the success rates of launches and the types of orbits achieved.



RESULTS

Matplotlib and Seaborn (EDA with Visualization)

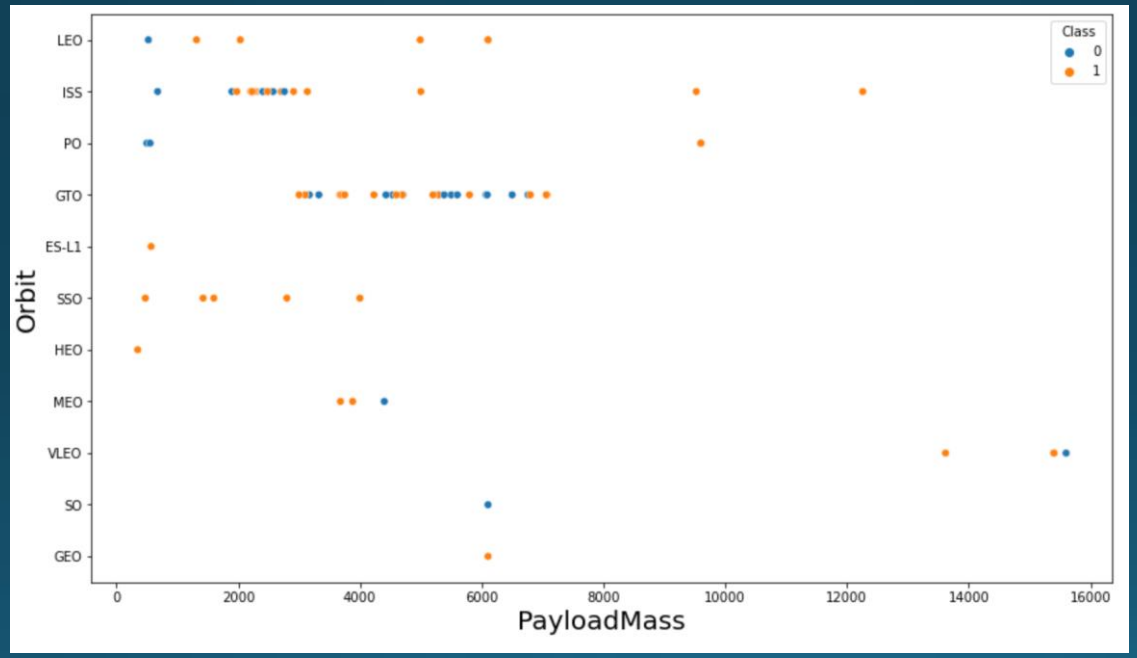
The association between the sequence of flights and the types of orbits utilized.



RESULTS

Matplotlib and Seaborn (EDA with Visualization)

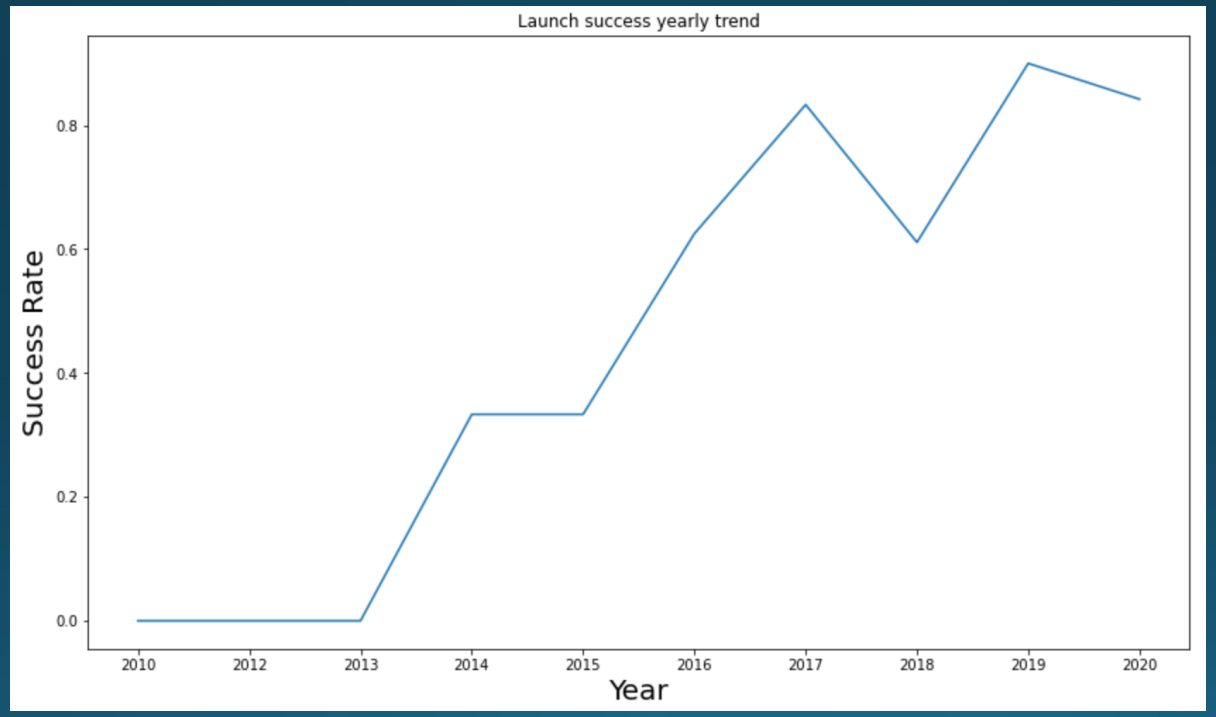
The interplay between the mass of payloads and the types of orbits they are sent into.



RESULTS

Matplotlib and Seaborn (EDA with Visualization)

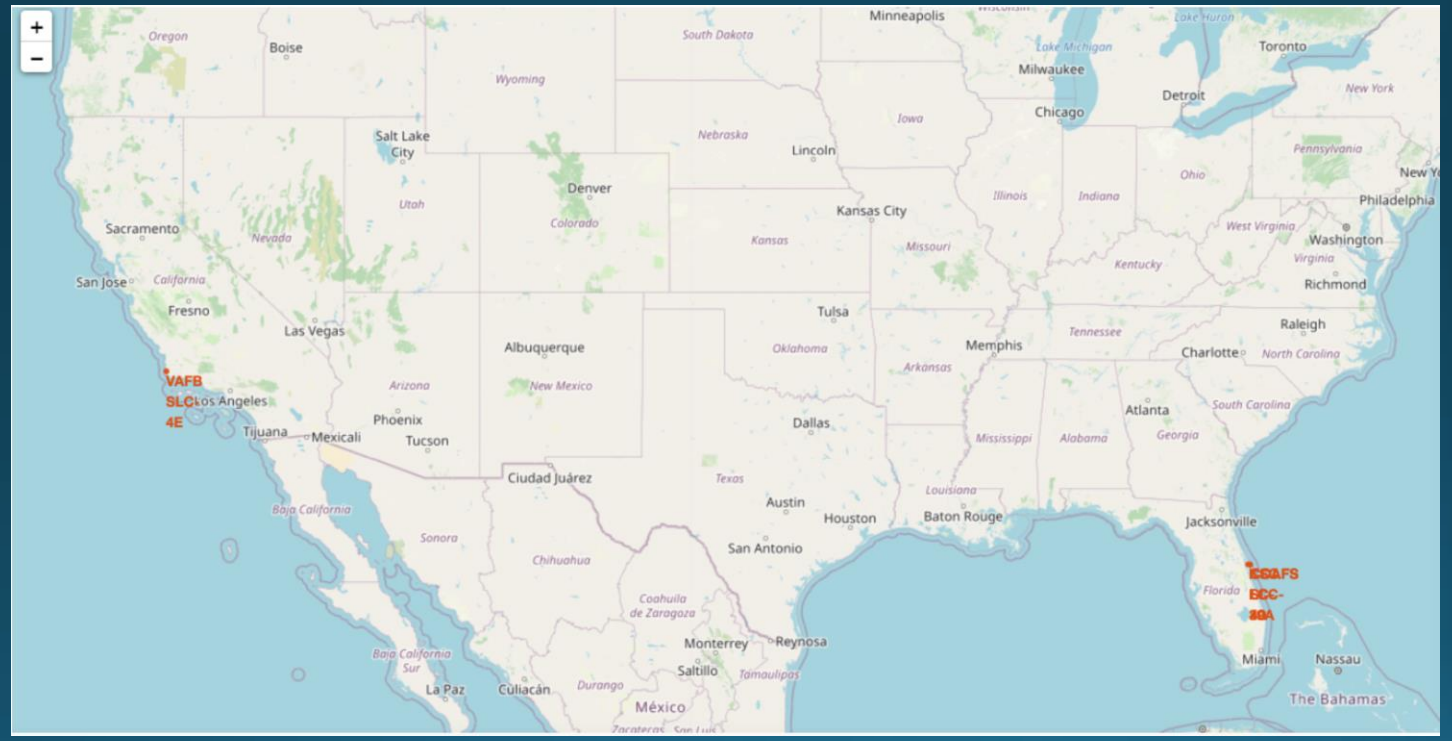
The annual pattern of launch successes over time.



RESULTS

Folium

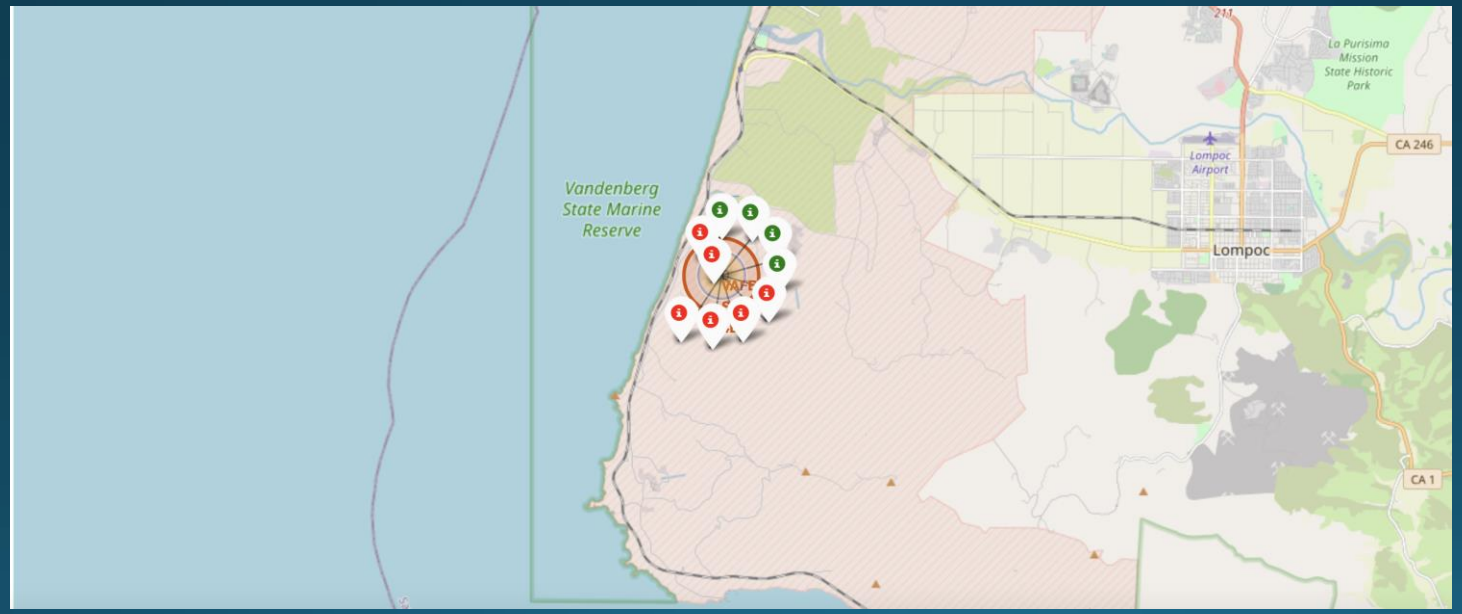
The geographical representation of all launch locations.



RESULTS

Folium

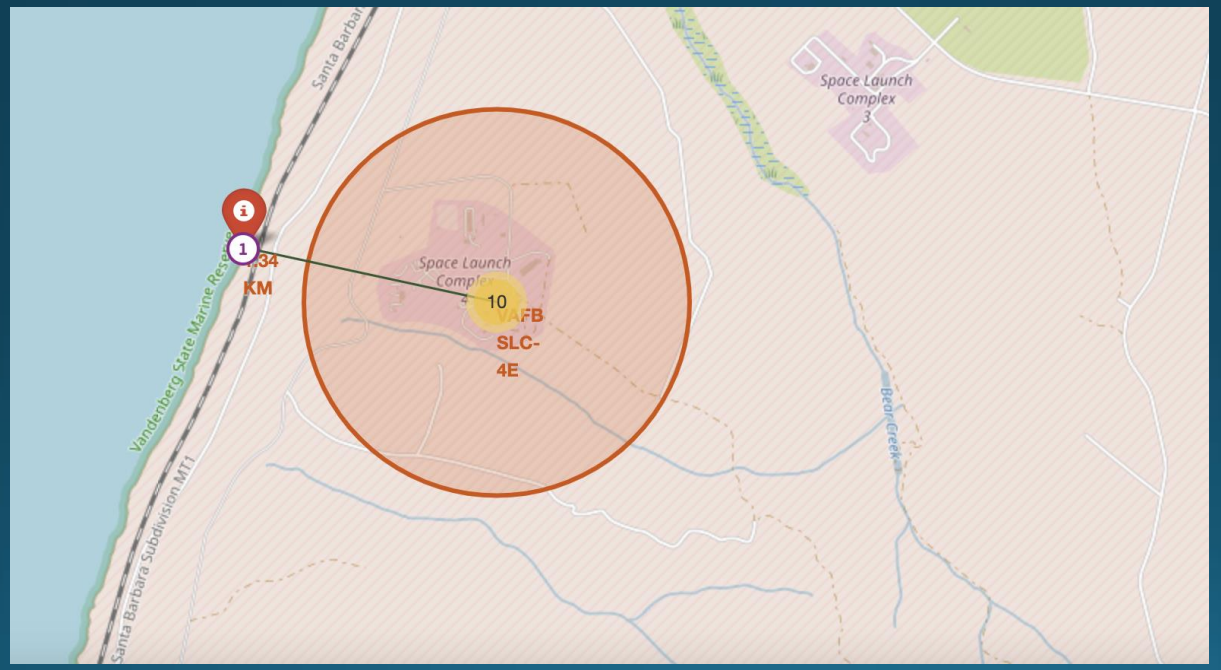
On the map displaying each launch site, successful and unsuccessful launches are marked with green and red pins respectively, with green indicating a launch success and red indicating a failure.



RESULTS

Folium

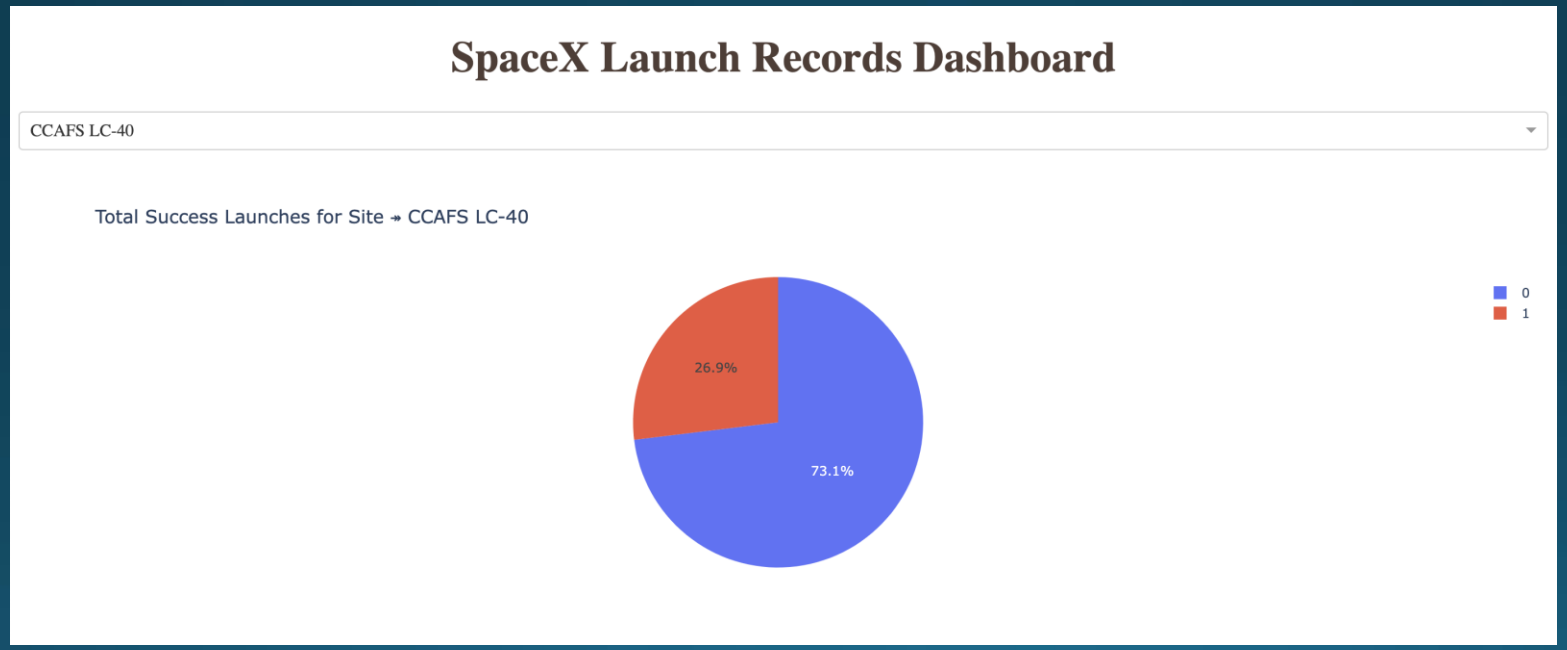
The visualization displays the distances from the VAFB SLC-4E launch site to nearby important infrastructure like cities, railways, and highways, highlighting the closest coastline.



RESULTS

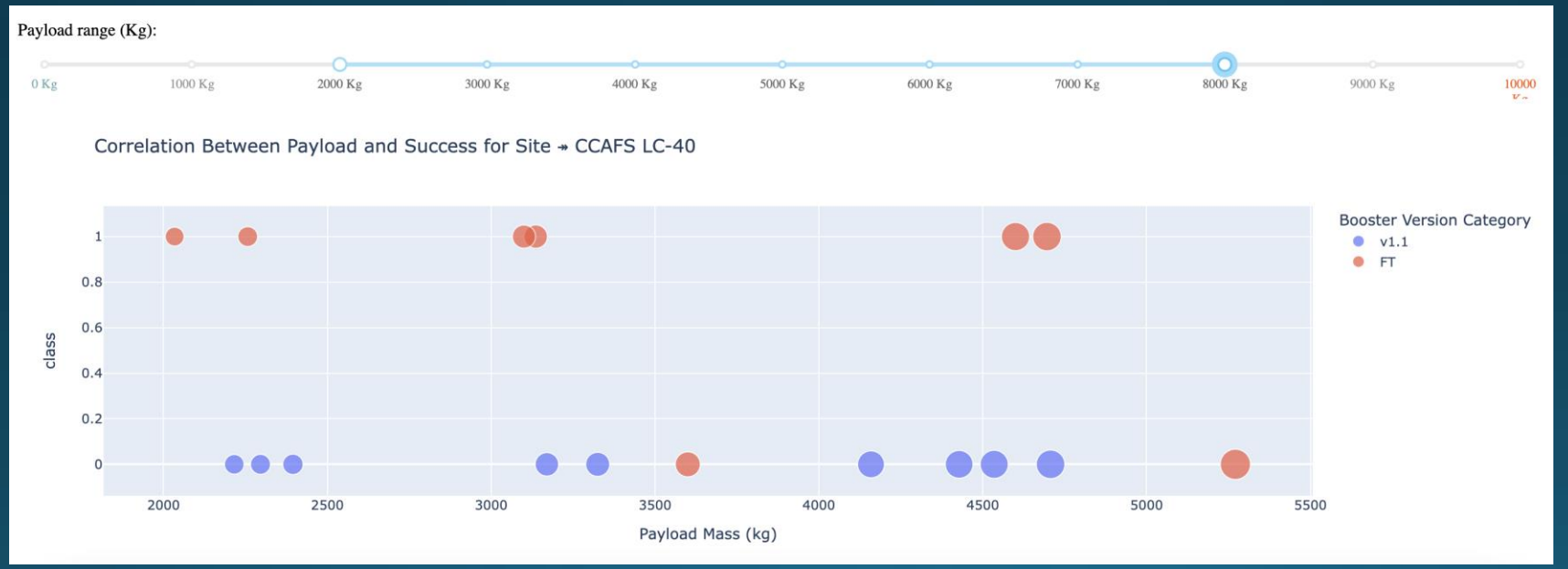
Dash

The displayed pie chart corresponds to the CCAFS LC-40 launch site, indicating that 73.1% of launches at this site did not achieve a successful landing.



RESULTS
Dash

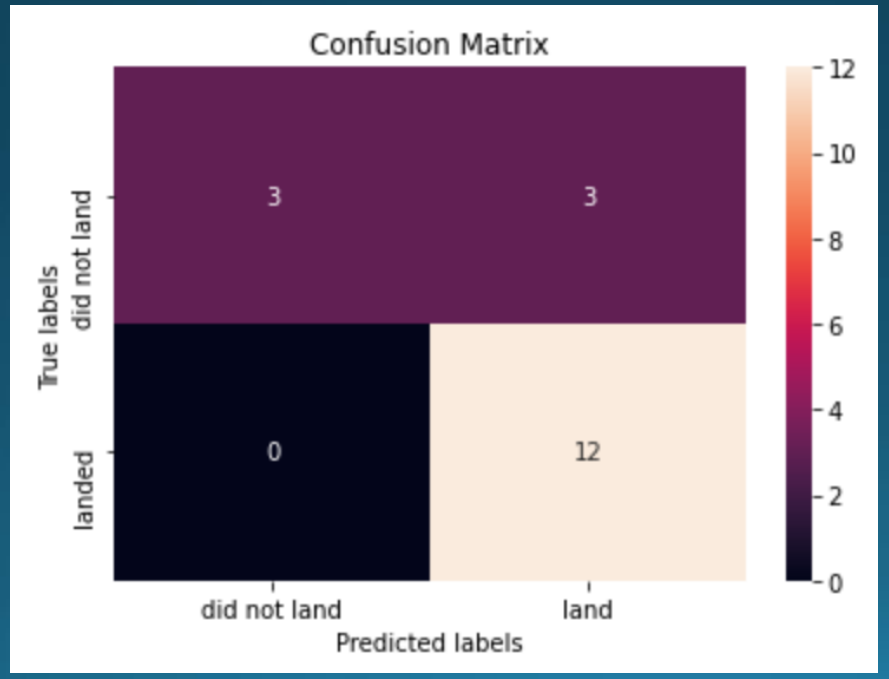
The image provided displays a scatterplot with the payload mass range from 2000kg to 8000kg. In this visualization, class 0 is indicative of launch failures and class 1 of launch successes.



RESULTS

Predictive Analysis

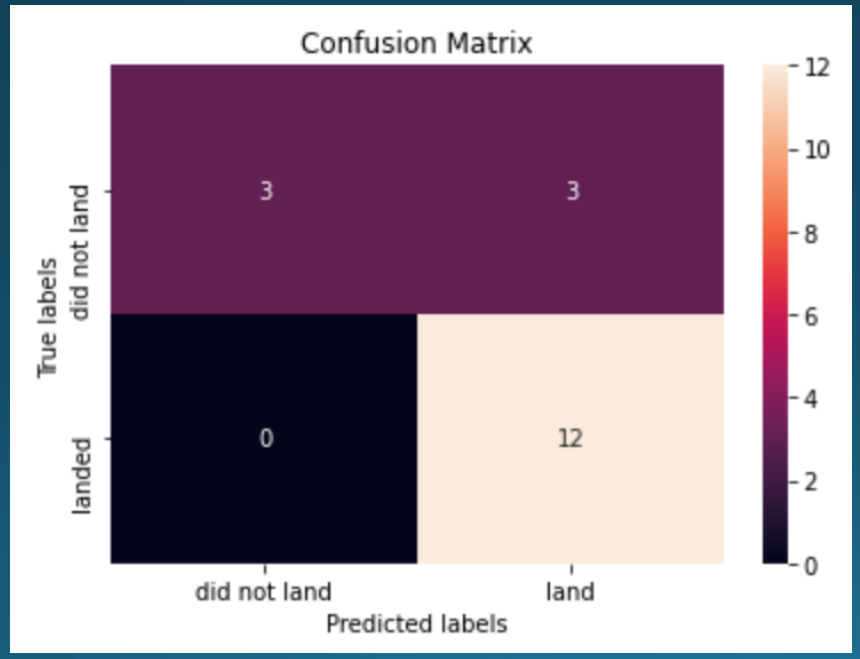
For logistic regression, the optimal score achieved through GridSearchCV is approximately 0.846, while the accuracy on the test set is about 0.833. The confusion matrix visualizes the true positive and false positive rates, indicating 12 correct predictions for landings, 3 incorrect non-landing predictions, and no false positives for landings.



RESULTS

Predictive Analysis

The Support Vector Machine model has achieved a top score of approximately 0.848 via GridSearchCV, with an accuracy of about 0.833 on the test dataset. The confusion matrix indicates 12 successful landing predictions and 3 misclassified as unsuccessful landings, with no instances of false positives for successful landings.

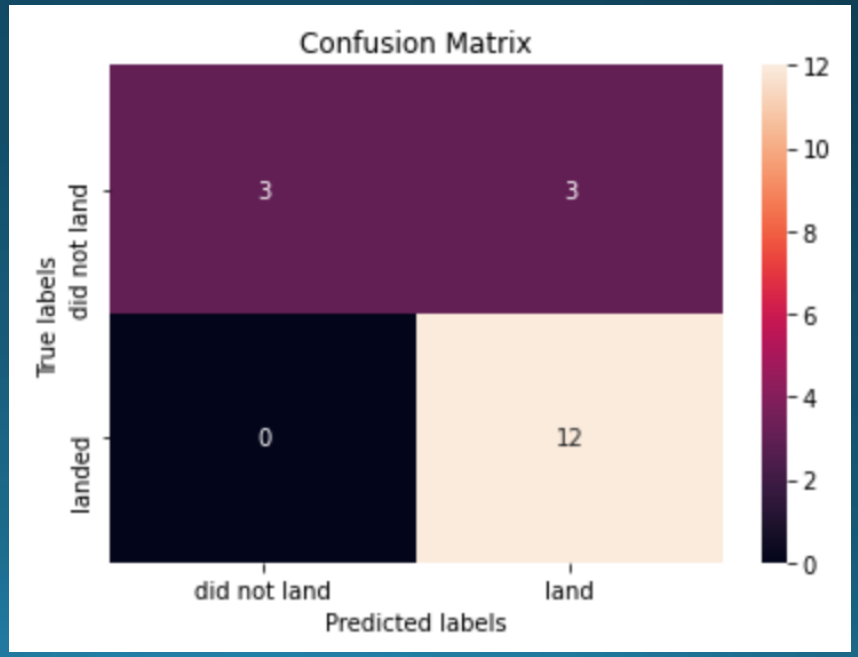


RESULTS

Predictive Analysis

The Decision Tree model has achieved a GridSearchCV best score of approximately 0.8892857142857142, with the model's accuracy on the test data being about 0.8333333333333334.

The associated confusion matrix shows that there were 12 correct predictions for landing and 3 incorrect predictions for non-landing, with zero false positives for landing.

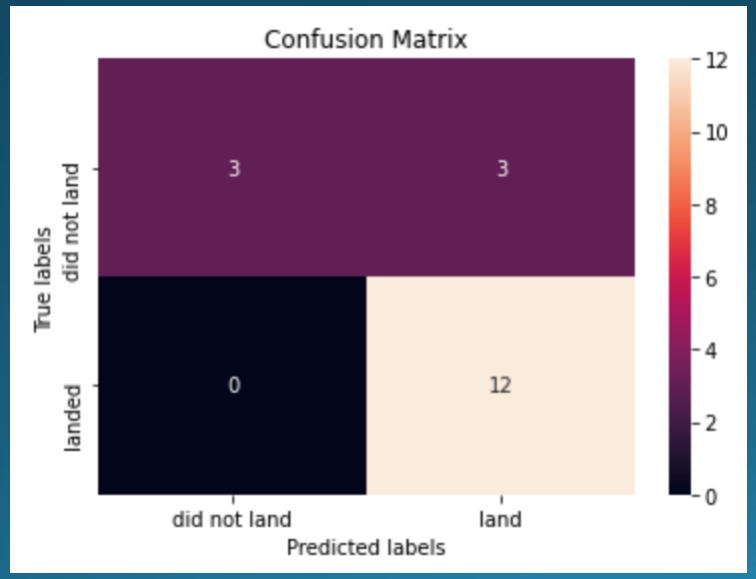


RESULTS

Predictive Analysis

The K Nearest Neighbors (KNN) model has secured the highest GridSearchCV score of approximately 0.8482142857142858 and achieved an accuracy of about 0.8333333333333334 on the test set.

The confusion matrix demonstrates 12 accurate predictions for successful landings and 3 misclassified predictions for non-landings, with no successful landings incorrectly predicted as failures.



RESULTS

Predictive Analysis

Upon comparative evaluation, the four machine learning models demonstrate identical accuracy and confusion matrix outcomes on the test set. They are ranked by their respective GridSearchCV best scores as follows:

- Decision tree with a score of 0.8892857142857142
- K nearest neighbors (KNN) with a score of 0.8482142857142858
- Support vector machine (SVM) with a score of 0.8482142857142856
- Logistic regression with a score of 0.8464285714285713

DISCUSSION

The visual data analysis suggests certain features may influence the outcome of the mission in various ways. For instance, heavier payloads appear to correlate with a higher success rate in orbits such as Polar, LEO, and ISS. However, with GTO orbits, it is less clear-cut as both successful and unsuccessful landings are observed.

As a result, it is understood that each feature can impact the mission result to some extent. While it is challenging to determine precisely how these features affect the outcome, machine learning algorithms can be applied to discern patterns from historical data to forecast the likelihood of mission success based on the available features.

CONCLUSION

The objective of this project is to forecast whether the initial stage of a Falcon 9 rocket will successfully land, thus assisting in cost estimation for each launch. Various aspects of a Falcon 9 launch, including payload weight and orbital trajectory, potentially influence mission results.

A suite of machine learning techniques has been applied to historical launch data to construct models capable of predicting Falcon 9 launch outcomes. Of the four machine learning models tested, the decision tree model emerged as the most effective.