IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

André Silva
30 December 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- In this capstone project, we aim to predict the success or failure of SpaceX Falcon 9 first stage landings. We will use a range of machine learning algorithms to make predictions and determine the success rate of success and determine the cost of a launch. This include Data Collection, Data Wrangling, Data Preprocessing, Data Exploratory Data Analysis, Data Visualization and Machine Learning Prediction.
- Our analysis suggests that certain characteristics of rocket launches may be related to the success or failure of the first stage landing. Based on our findings, we conclude that the Decision Tree algorithm may be the best choice for predicting the outcome of these landings.

Introduction

- The primary objective of this capstone project is to develop a prediction model that can accurately determine the success or failure of Falcon 9 first stage landings. Reusability of the first stage is a key factor in the cost of SpaceX's rocket launches, which are advertised as being significantly cheaper than those of other providers. By predicting the outcome of first stage landings, we can potentially estimate the cost of a rocket launch and use this information to compete with SpaceX in the bidding process for future launches.
- The main questions is if certain characteristics of rocket launches may be related to the success or failure of SpaceX Falcon 9 first stage landings?

Section 1

Methodology

Methodology

Executive Summary

- To gather the data for this project, we employed API requests to SpaceX and web scraping of launch data from a Wikipedia page.
- We then used Python's pandas library to clean and transform the data.
- For our exploratory data analysis, we utilized a range of visualization tools such as matplotlib, seaborn, and folium, as well as SQL queries and interactive visualizations created with Plotly Dash.
- For the predictive analysis, we applied four different machine learning classification models: logistic regression, support vector machines, k-nearest neighbor, and decision tree classifier.
- We trained, tuned, and evaluated each model to determine the most effective one for this problem.

Data Collection – SpaceX API

- **Data collection methodology**:

  1. Request and parse the SpaceX launch data using the GET request
  2. Normalize JSON response into a dataframe
  3. Extract useful columns
  4. Filter the dataframe to only include Falcon 9 launches
  5. Dealing with Missing Values
  6. Export to CSV file

- https://github.com/AndreSilva101/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-spacex%20data-collection%20API.ipynb

Data Collection - Scraping

Web scraping

- **Data collection methodology**:

    1. Request the Falcon9 Launch Wiki page from its URL
    2. Extract all column/variable names from the HTML table header
    3. Parsing the launch HTML tables and create a data frame
    4. Export to CSV

- https://github.com/AndreSilva101/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs%20WEBSCRAPING.ipynb

Data Wrangling

- Perform data wrangling

1. Calculate the number of launches on each site
2. Calculate the number and occurrence of each orbit
3. Calculate the number and occurrence of mission outcome per orbit type
4. Create a landing outcome label from Outcome column
5. Export to CSV

- https://github.com/AndreSilva101/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/labs-jupyter-spacex%20DATA_WRANGLING%20jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

Exploratory data analysis (EDA) using visualization

- Scatter plots: These plots were used to explore the relationships between different pairs of variables, such as the flight number and launch site, payload and launch site, flight number and orbit type, and payload and orbit type.
- Bar charts: These charts were used to compare values between different groups by displaying a bar for each category on the x-axis and the corresponding value on the y-axis. We used bar charts to compare the success rate for different orbit types.
- Line charts: These charts are useful for visualizing trends in data over time, and we used a line chart to show the success rate over a number of years.

- https://github.com/AndreSilva101/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs%20EDA%20DATA%20VISUAL.ipynb.jupyterlite.ipynb

EDA with SQL

Exploratory data analysis (EDA) using SQL

- Displaying the names of the unique launch sites used in space missions.
- Display 5 records where launch sites begin with the string 'KSC' .
- Display the total payload mass carried by boosters launched by NASA (CRS).
- Display average payload mass carried by booster version F9 v1.1.
- List the date where the succesful landing outcome in drone ship was acheived.
- List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000.
- List the total number of successful and failure mission outcomes.
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery.
- List the records which will display the month names, successful landing outcomes in ground pad ,booster versions, launch_site for the months in year 2017.
- Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

- https://github.com/AndreSilva101/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs%20%20EDA-SQL%20%20edx_sqllite.ipynb

Methodology

Executive Summary

- **Perform interactive visual analytics using Folium and Plotly Dash**
- We used Folium to create and add objects to a map. Marker objects were used to show all launch sites on the map, as well as the successful and failed launches for each site. Line objects were used to calculate the distances between a launch site and its proximities.
    - Are launch sites in close proximity to railways? Yes
    - Are launch sites in close proximity to highways? Yes
    - Are launch sites in close proximity to coastline? Yes
    - Do launch sites keep certain distance away from cities? Yes

- https://github.com/AndreSilva101/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/lab_jupyter%20Launch%20SITE%20location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

- A pie chart displays the successful launches by site, allowing you to see the distribution of landing outcomes among all launch sites or the success rate of launches at individual sites.
- A scatter chart shows the relationship between landing outcomes and the payload mass of various boosters. This chart can be filtered by site and payload mass and is useful for understanding how different variables impact landing outcomes.

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
- You need present your model development process using key phrases and flowchart
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Results

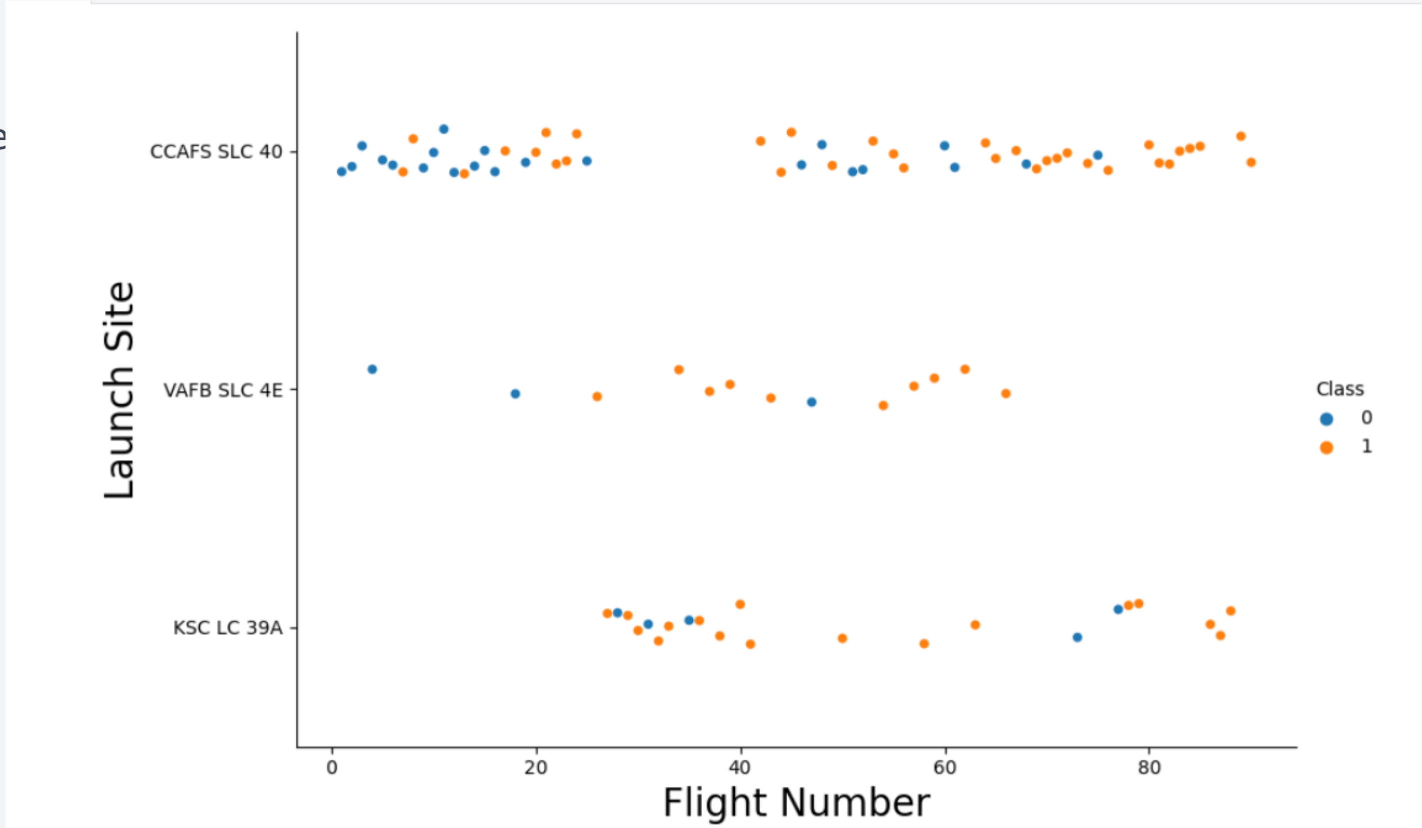
- The success rate of Falcon 9 landings, as determined through data analysis, is 66.66%.
- Predictive analysis using the Decision Tree algorithm yielded the best classification results, with an accuracy of 94%.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- The graph illustrates that the success rate improved as the number of flights increased.
- The orange dots represent successful launches and the blue dot represents unsuccessful launches. It appears that there was a rise in successful flights after the 40th launch.

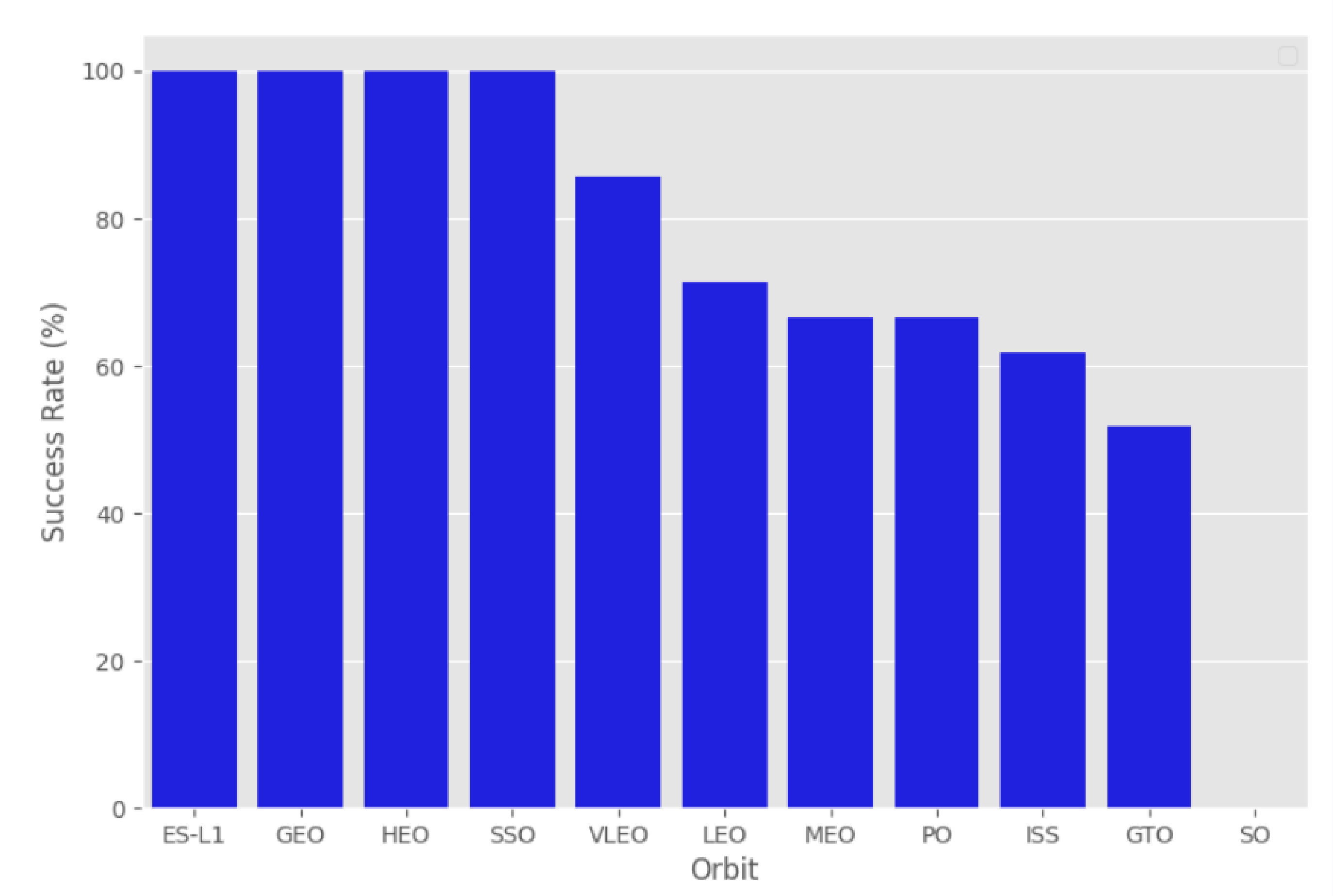


Payload vs. Launch Site

- On the graph, orange dots represent successful launches and blue dots represent unsuccessful launches. For the VAFB-SLC launch site, there are no rockets launched for heavy payload mass.
- The relationship between payload mass and launch site appears to be weak, and thus it is not useful for making decisions

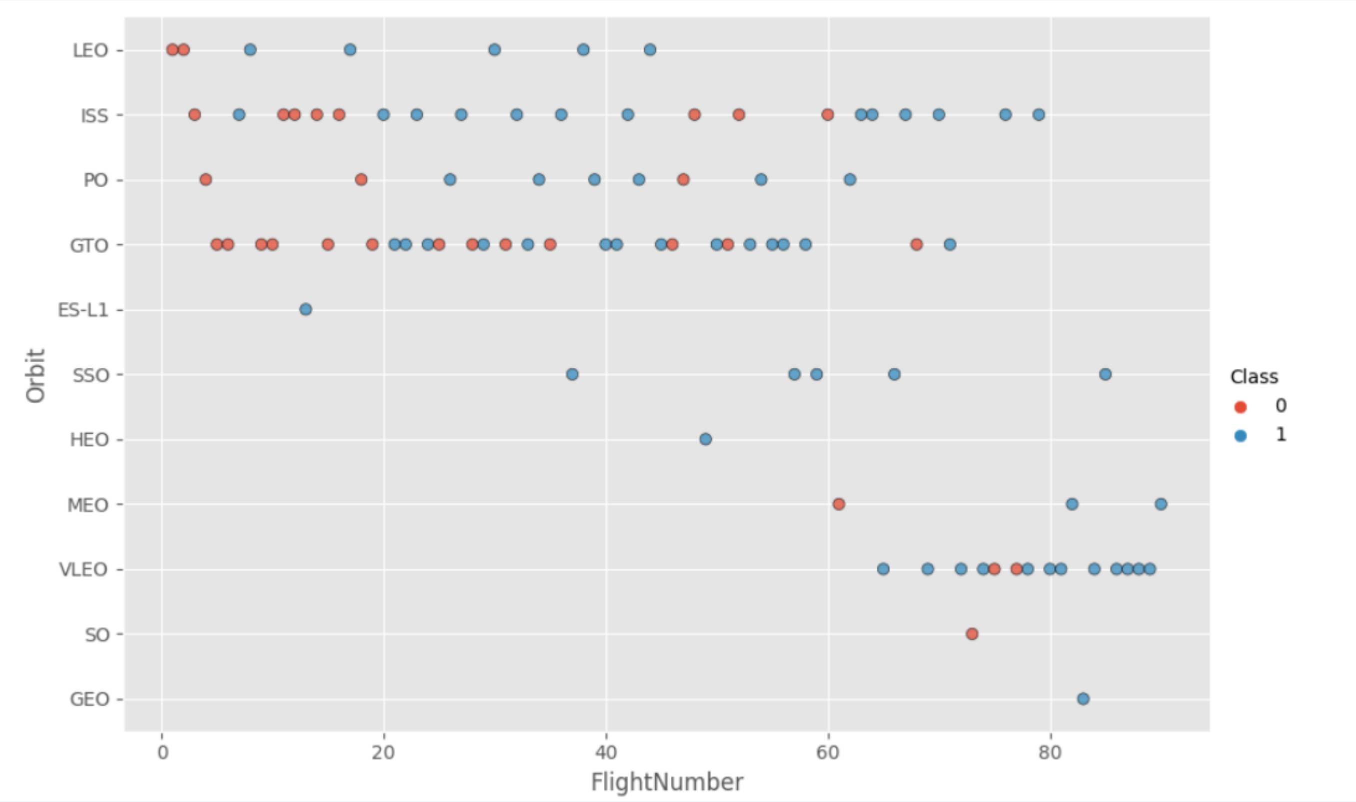
Success Rate vs. Orbit Type

- SO orbit did not have any successful launches with a 0% success rate.
- Orbits SSO, HEO, GEO, and ES-L1 have 100% success rates.



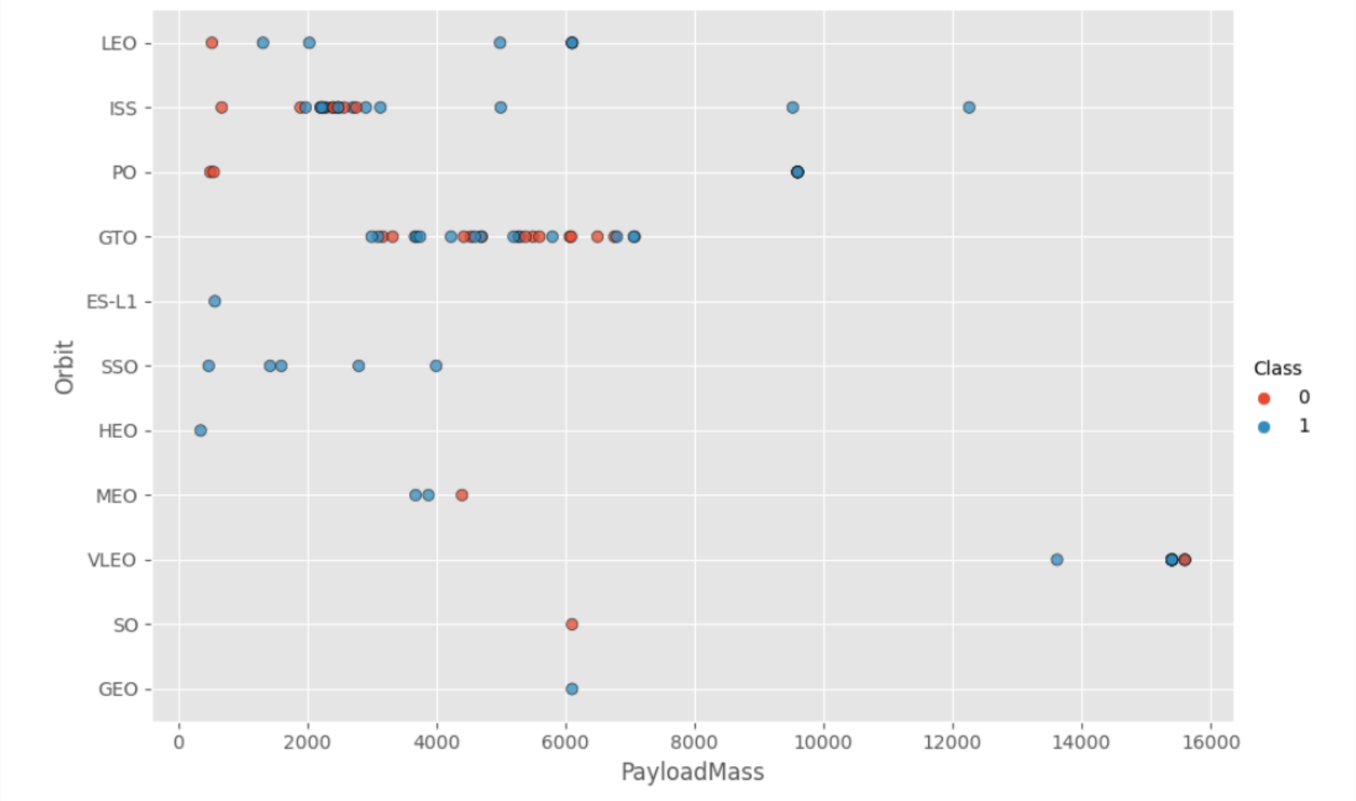
Flight Number vs. Orbit Type

- In the LEO orbit, success is linked with an increase in the number of flights.
- The SSO orbit has a 100% success rate, but with fewer flights than the other orbits.
- Flights numbered above 40 have a higher success rate than flights numbered between 0 and 40.
- There is no connection between flight number and success in the GTO orbit.



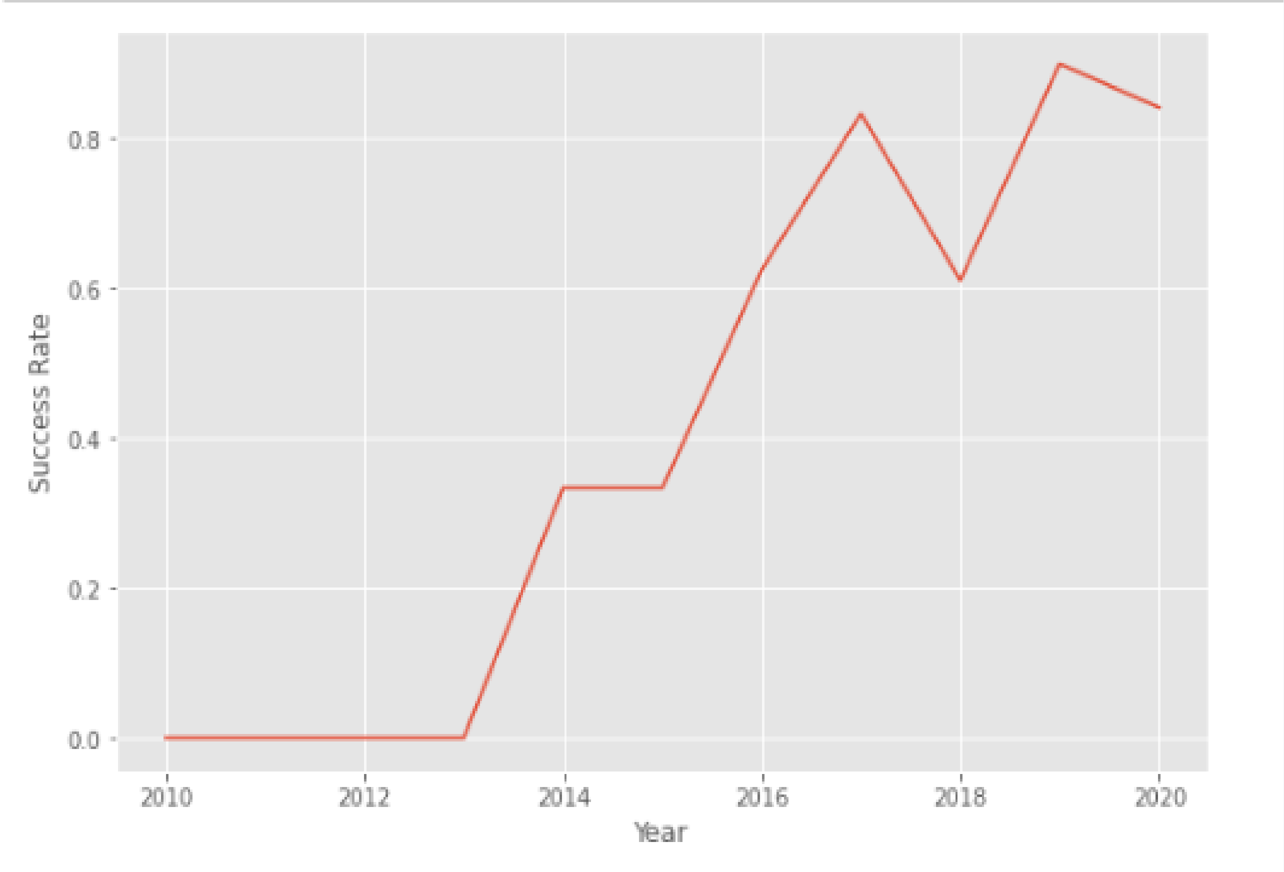
Payload vs. Orbit Type

- As payload mass increases, success rate improves in the PO, SSO, LEO, and ISS orbits. There is no clear relationship between orbit type and payload mass for the GTO orbit, as both successful and failed launches are present in roughly equal numbers.



Launch Success Yearly Trend

- The chart shows that the success rate of landings generally increases over time, but there are dips in 2018 and 2020. This indicates that the success rate is not consistently improving



All Launch Site Names

- DISTINCT clause was used to return only the unique rows from the *launch_site* column.

Display the names of the unique launch sites in the space mission

In [7]:
```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

Out[7]:

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

Launch Site Names Begin with 'KSC'

- The LIMIT 5, display the top five results.
- LIKE clauses were used to display only where the *launch_site* name starts with 'KSC'

Task 2

Display 5 records where launch sites begin with the string 'KSC'

In [9]:

```
%sql SELECT * FROM 'SPACEXTBL' WHERE Launch_Site LIKE 'KSC%' LIMIT 5
```

```
 * sqlite:///my_data1.db
Done.
```

Out[9]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
19-02-2017	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
16-03-2017	06:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
30-03-2017	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
01-05-2017	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
15-05-2017	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

- To calculate the total payload carried by boosters from NASA from the *payload_mass__kg* column, we used the SUM() function

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

In [12]:
```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as PM_KG_TOTAL, Customer FROM 'SPACEXTBL' WHERE Customer = 'NASA (CRS)'
```

```
 * sqlite:///my_data1.db
Done.
```

Out[12]:

PM_KG_TOTAL	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

- The AVG() function was used to the calculate the average payload the average payload mass carried by booster version F9 v1.1
- The WHERE clause was used to filter results so that the calculations were only performed on *booster_versions* only if they were named "F9 v1.1"

Task 4

Display average payload mass carried by booster version F9 v1.1

In [13]:
```
%sql SELECT AVG(PAYLOAD_MASS__KG_) as PM_KG_AVG FROM 'SPACEXTBL' WHERE Booster_Version LIKE 'F9 v1.1%'
```
```
 * sqlite:///my_data1.db
Done.
```
Out[13]:

PM_KG_AVG
2534.6666666666665

First Successful Ground Landing Date

- MIN(DATE) : used to find the date of the first successful landing outcome
- WHERE clause filtered to match only when the *'landing*_outcome' column is 'Success (drone ship)'

Task 5

List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

In [21]:

```
%%sql
SELECT min(DATE) AS "First successful landing outcome in drone ship" FROM SPACEXTBL WHERE "Landing _Outcome" = 'Success (drone ship)';
```

```
 * sqlite:///my_data1.db
Done.
```

Out[21]:

First successful landing outcome in drone ship
06-05-2016

Successful Drone Ship Landing with Payload between 4000 and 6000

- DISTINCT clause was used to return only the unique rows from the *launch_site* column.
- WHERE clause filtered the results to include only boosters which successfully landed on ground pad.
- BETWEEN clause was used to retrieve only those results of payload mass greater than 4000 but less than 6000.
- %sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Mission_Outcome = 'Success' and "Landing _Outcome" = 'Success (ground pad)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1

Total Number of Successful and Failure Mission Outcomes

- COUNT() is used to count the number of occurences and with the help of the GROUPBY clause applied to the Landing _Outcome' column, outputs a list of the total number of successful, failure and all mission.
- Success: 61
- Failure: 40
- (All): 101

Task 7

List the total number of successful and failure mission outcomes

```
%%sql
SELECT 'Success' AS "Outcome", count(*) AS "Count" FROM SPACEXTBL WHERE "Landing _Outcome" LIKE 'Success%'
UNION ALL
SELECT 'Failure' AS "Outcome", count(*) AS "Count" FROM SPACEXTBL WHERE "Landing _Outcome" NOT LIKE 'Success%'
UNION ALL
SELECT '(All)' AS "Outcome", count(*) AS "Count" FROM SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

Outcome	Count
Success	61
Failure	40
(All)	101

Boosters Carried Maximum Payload

- MAX() function was used in a subquery to retrieve a list of boosters which have carried the maximum payload mass.

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [26]:
```
%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

Out[26]:

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

2017 Launch Records

- substr(Date, 4, 2) was used to retrieve the month of the column "Date” .
- substr(Date, 7, 4) was used to retrieve the year of the column "Date” = 2017 Launch Records.
- WHERE clause filtered the results to include only boosters which successfully landed on ground pad.
- %sql SELECT substr(Date, 4, 2) as month, BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE substr(Date,7,4) = '2017' AND "Landing _Outcome" = 'Success (ground pad)';

month	Booster_Version	Launch_Site
02	F9 FT B1031.1	KSC LC-39A
05	F9 FT B1032.1	KSC LC-39A
06	F9 FT B1035.1	KSC LC-39A
08	F9 B4 B1039.1	KSC LC-39A
09	F9 B4 B1040.1	KSC LC-39A
12	F9 FT B1035.2	CCAFS SLC-40

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

- BETWEEN clause was used to retrieve only those results of DATE BETWEEN '04-06-2010' and '20-03-2017' .
- Count(*) was used to count the Landing _Outcome occurrences, with GROUP BY Landing _Outcome.
- %sql SELECT [Landing _Outcome], COUNT(*) AS "Count" FROM SPACEXTBL WHERE DATE BETWEEN '04-06-2010' and '20-03-2017' GROUP BY [Landing _Outcome] ORDER BY Count DESC ;

Landing _Outcome	Count
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Folium Map : US SpaceX Launch Sites Locations

- Yellow markers are the locations of all the SpaceX launch sites are situated in the US.

Folium Map : Launch display marker

- When we zoom in on a launch site, we can click on the launch site which will display marker clusters of successful landings (green) or failed landing (red).

Folium Map : Launch Site

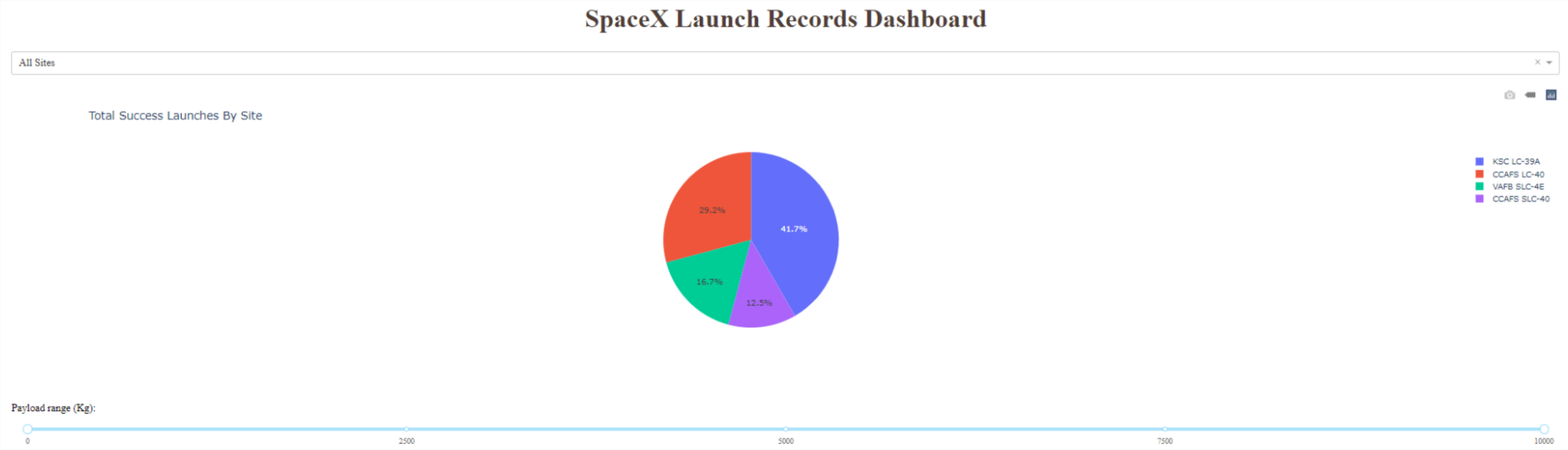
- Shows the selected launch site is close to a highway for transportation of personnel and equipment. The launch site is also close to the coastlines for launch failure testing.
- The launch sites also maintain a certain distance from the cities.

Section 4

Build a Dashboard with Plotly Dash

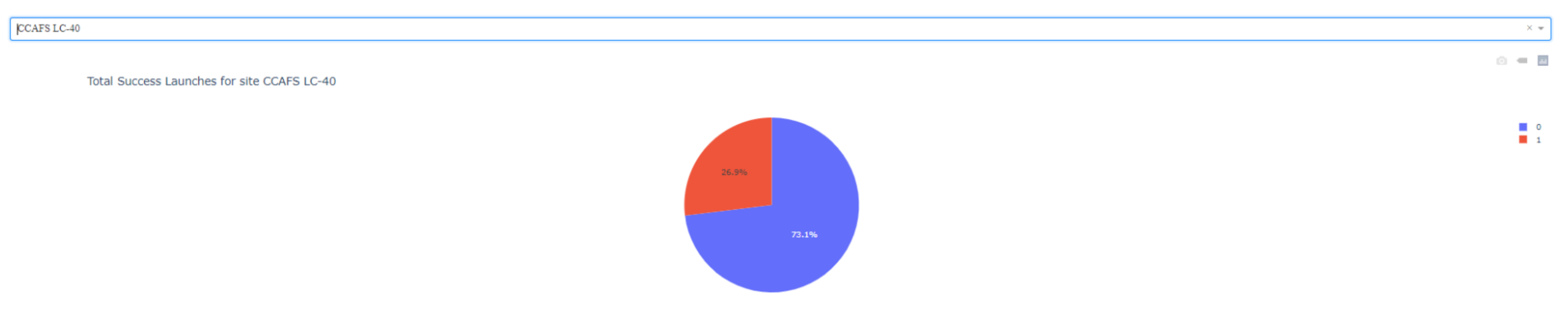
Total Successful Launches By Site

- The KSC LC-39A Launch site has the most successful launches.



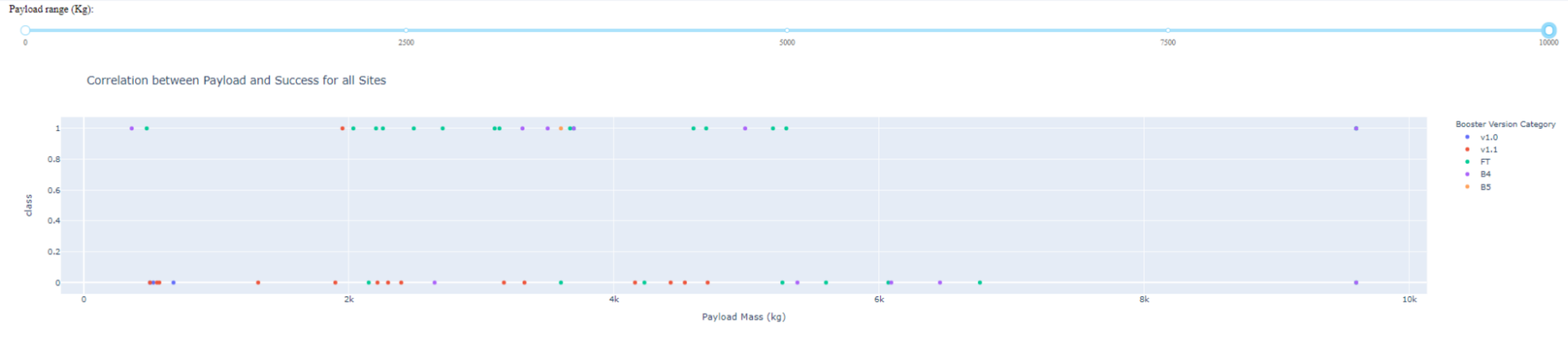
Total Success Launches for site CCAFS LC-40

- The CCAFS LC-40 has the success rate with 26.9%



Dashboard: Booster Versions V1.0, V1.1

- Success rate for Booster versions FT, B4 and B5 is better in the payload range to 10000kg

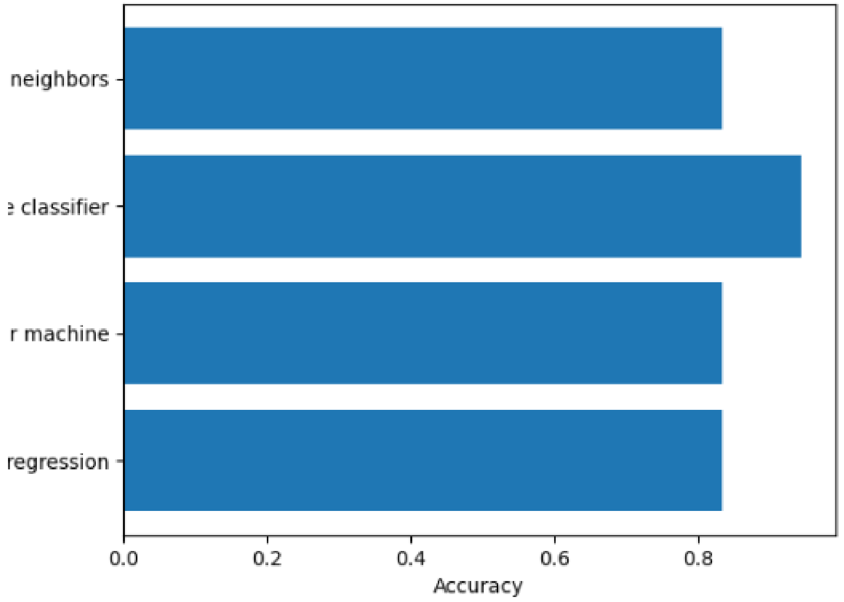


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The Decision Tree classifier had the best accuracy at 94%.



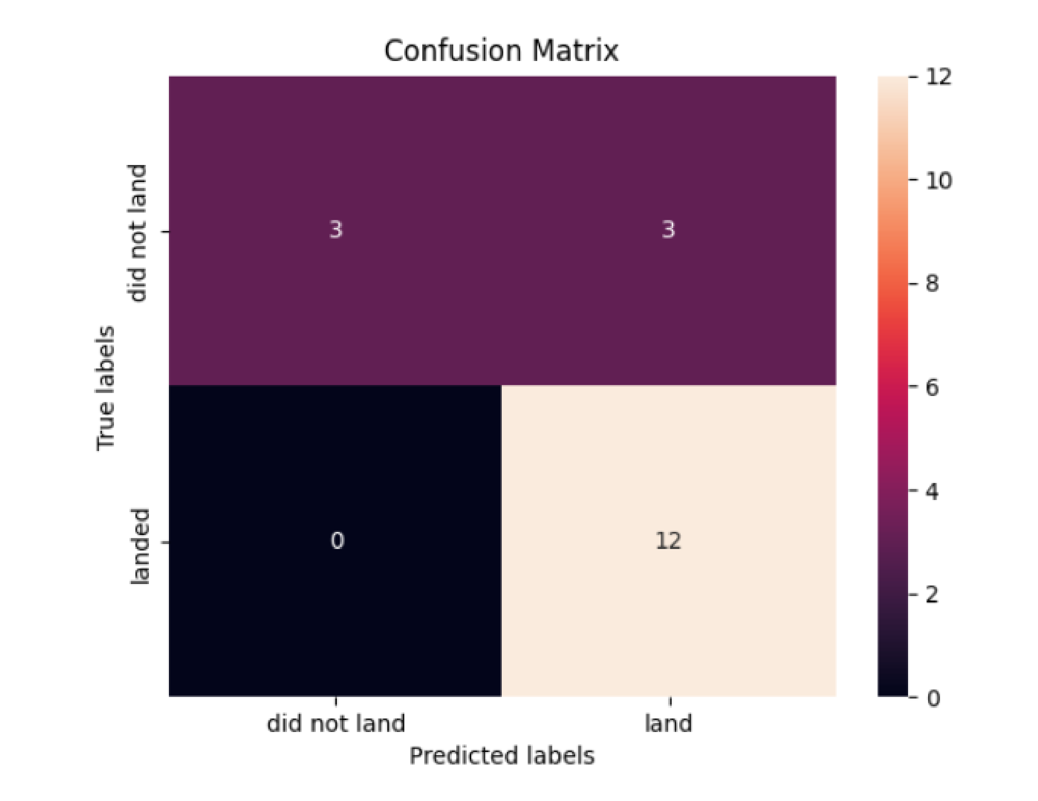
[34]:
```
results_df = {'method': method, 'accuracy': accuracy}

frame = pd.DataFrame(results_df)
frame
```

[34]:

	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333

Confusion Matrix

- The model predicted 12 successful landings when the true label was successful (true positive), and 3 unsuccessful landings when the true label was unsuccessful (True Negative).
- However, the model also predicted 3 successful landings when the true label was unsuccessful landing (False Positive).
- Overall, the model tends to predict successful landings.



Conclusions

- The analysis showed that there is a positive relationship between the number of flights and success rate, as success rate has improved over time.
- Some orbits, such as SSO, HEO, GEO, and ES-L1, had the highest success rates for launches.
- The success rate may also be influenced by payload mass, with lighter payloads generally having higher success rates than heavier payloads.
- The launch sites are located near transportation infrastructure, such as highways and railways, to facilitate the movement of personnel and cargo, but they are also situated far from urban areas for safety reasons.
- Based on the results of the analysis, the best predictive model for this dataset is the Decision Tree Classifier, which had an accuracy of 94%.

Appendix

- GitHub Repository Data-Science-and-Machine-Learning-Capstone-Project:
- https://github.com/AndreSilva101/Data-Science-and-Machine-Learning-Capstone-Project

Thank you!