IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Breno Nunes
06/07/2024

Outline

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

Executive Summary

Data was collected from the public SpaceX API and the SpaceX Wikipedia page. A label column named 'class' was created to classify successful landings. Data exploration was performed using SQL, visualizations, Folium maps, and dashboards. Relevant columns were gathered to be used as features. All categorical variables were converted to binary using one-hot encoding. The data was standardized, and GridSearchCV was used to find the best parameters for machine learning models. The accuracy scores of all models were visualized.

Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K-Nearest Neighbors. All produced similar results with an accuracy rate of about 83.33%. All models overpredicted successful landings. More data is needed for better model determination and accuracy.

Introduction

Background and context

- The Commercial Space Age has arrived
- SpaceX offers the best pricing ($62 million vs. $165 million USD)
- This is largely due to their ability to recover part of the rocket (Stage 1)
- Space Y aims to compete with SpaceX

Problems you want to find answers

- Space Y has tasked us with training a machine learning model to predict successful Stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection

The data collection process involved a combination of API requests from SpaceX's public API and web scraping data from a table on SpaceX's Wikipedia page. The next slide will present the flowchart for data collection from the API, and the following slide will show the flowchart for data collection from web scraping.

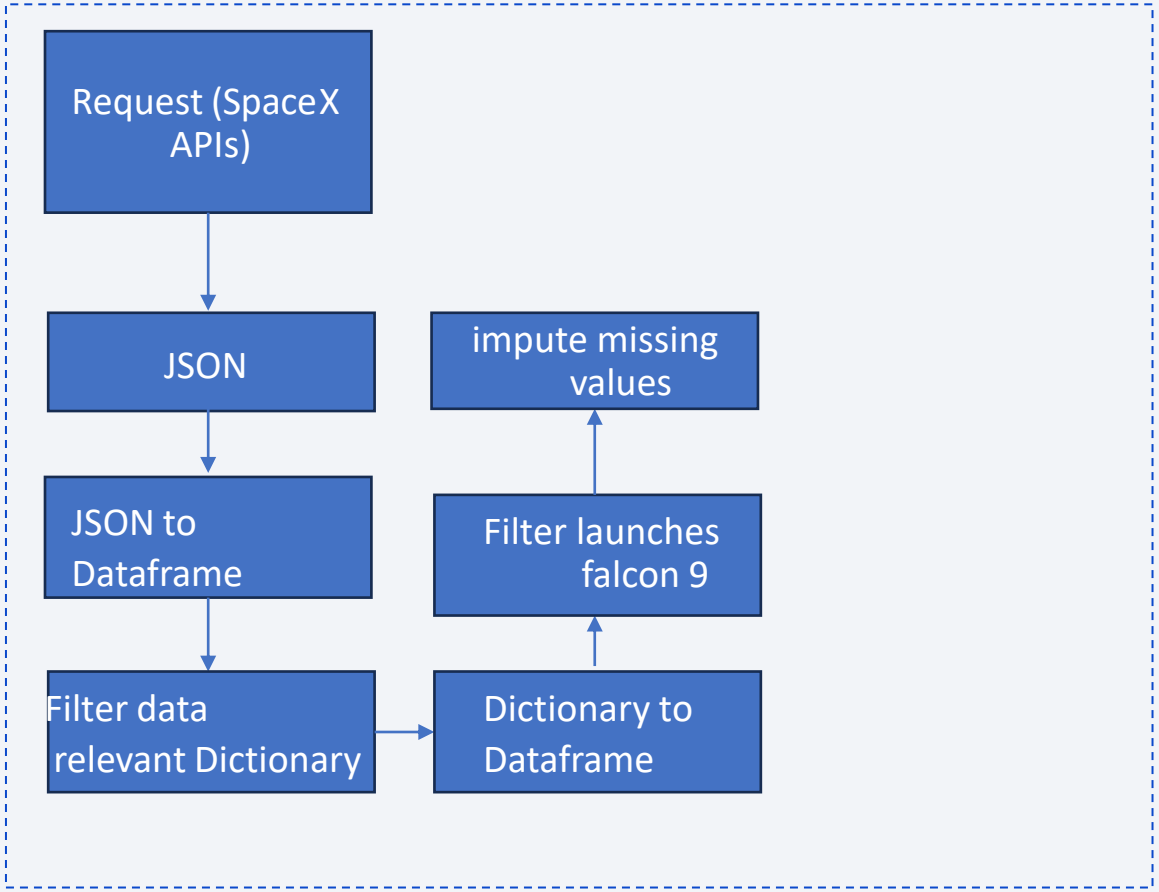
SpaceX API Data Columns: FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Web Scrape Data Columns: Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

- Present your data collection with SpaceX REST calls using key phrases and flowcharts

- Add the GitHub URL: https://github.com/bsensix/IBM_Data_Science_Professional_Certification/blob/main/Module_1/data_collection_api%20.ipynb

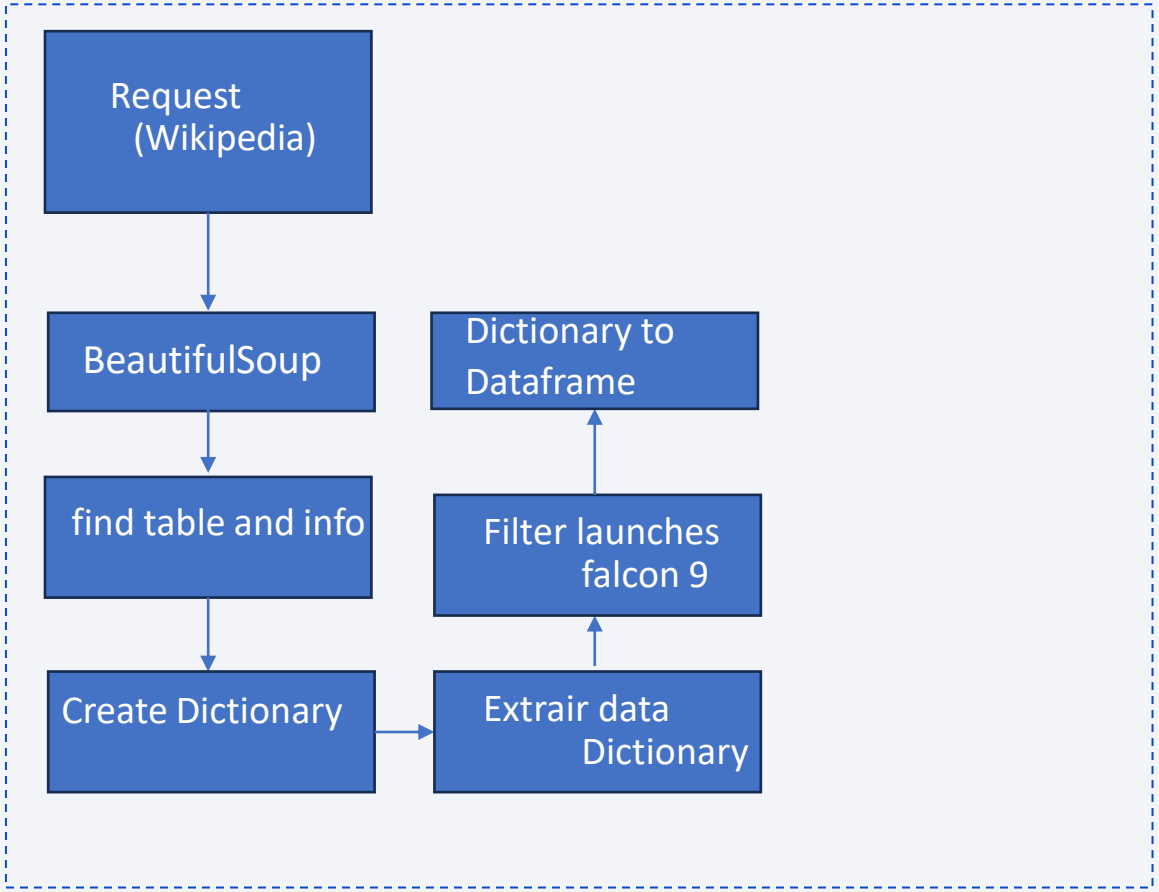


Data Collection - Scraping

- Present your web scraping process using key phrases and flowcharts

- Add the GitHub URL:

https://github.com/bsensix/IBM_Data_Science_Professional_Certification/blob/main/Module_1/data_collection_web_scraping.ipynb



Data Wrangling

- Describe how data were processed

Create a training label for landing outcomes where successful = 1 and failure = 0. The Outcome column has two components: 'Mission Outcome' and 'Landing Location'. Add a new training label column named 'class' with a value of 1 if the 'Mission Outcome' is True, and 0 otherwise. The value mapping is as follows:

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

- Add the GitHub URL: https://github.com/bsensix/IBM_Data_Science_Professional_Certification/blob/main/Module_1/data_wrangling.ipynb

EDA with Data Visualization

Summarize what charts were plotted and why you used those charts

Exploratory Data Analysis was performed on the variables Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. The plots used include: Flight Number vs. Payload MassFlight Number vs. Launch SitePayload Mass vs. Launch SiteOrbit vs. Success RateFlight Number vs. OrbitPayload vs. OrbitSuccess Yearly Trend.

Scatter plots, line charts, and bar plots were used to compare relationships between variables to determine if a relationship exists, which could then be used in training the machine learning model.

- Add the GitHub URL: https://github.com/bsensix/IBM_Data_Science_Professional_Certification/blob/main/Module_2/EDA_Visualization.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

The dataset was loaded into an IBM DB2 Database. Queries were performed using SQL Python integration to gain a better understanding of the dataset.

Queries were made to retrieve information about launch site names, mission outcomes, various payload sizes of customers, booster versions, and landing outcomes.

- Add the GitHub URL: https://github.com/bsensix/IBM_Data_Science_Professional_Certification/blob/main/Module_2/EDA_SQL.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map

Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations such as Railways, Highways, Coasts, and Cities. This helps us understand why launch sites are located in specific areas and visualizes successful landings relative to their locations.

- Add the GitHub URL:

  https://github.com/bsensix/IBM_Data_Science_Professional_Certification/blob/main/Module_3/interactive_visual_analytics_folium.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard

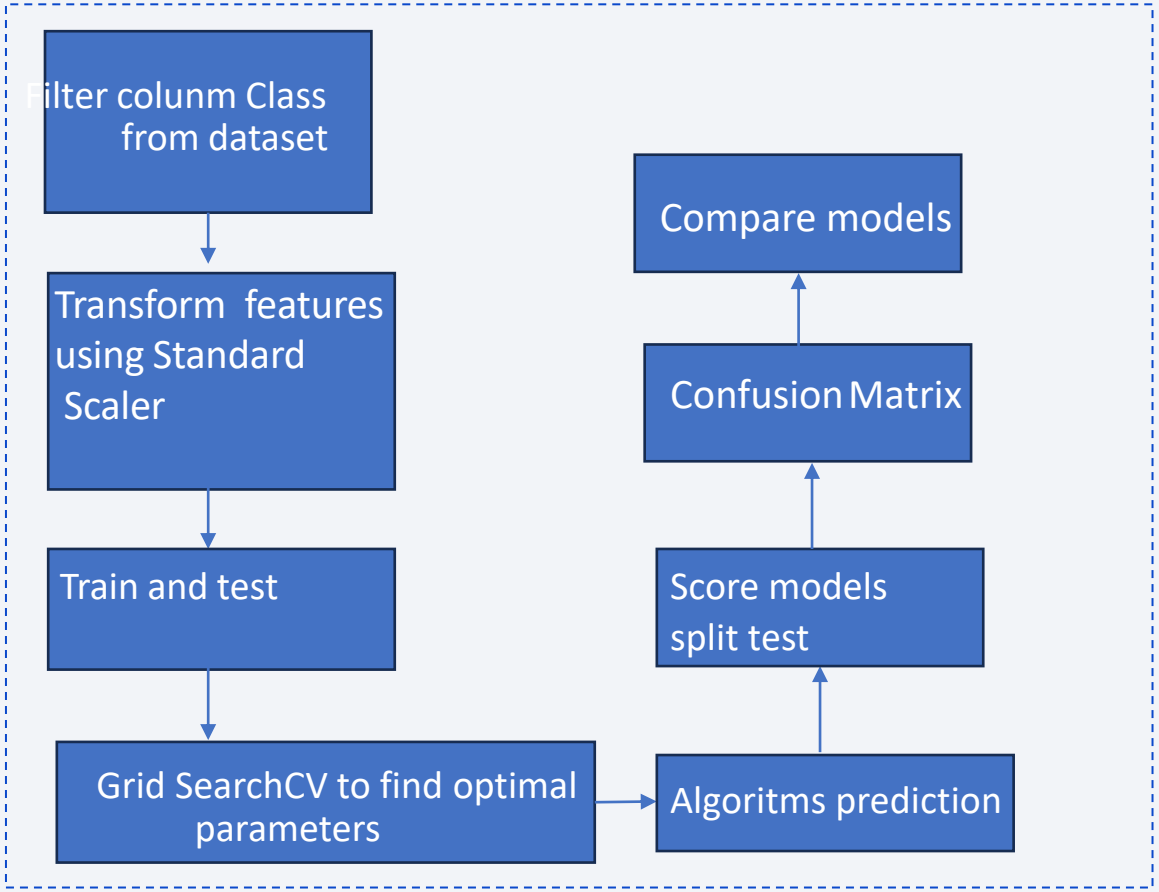
The dashboard includes a pie chart and a scatter plot. The pie chart can display the distribution of successful landings across all launch sites or the success rates of individual launch sites. The scatter plot takes two inputs: all sites or an individual site, and payload mass on a slider ranging from 0 to 10,000 kg. The pie chart is used to visualize the success rate of launch sites. The scatter plot helps us see how success varies across launch sites, payload mass, and booster version categories.

Add the GitHub URL:

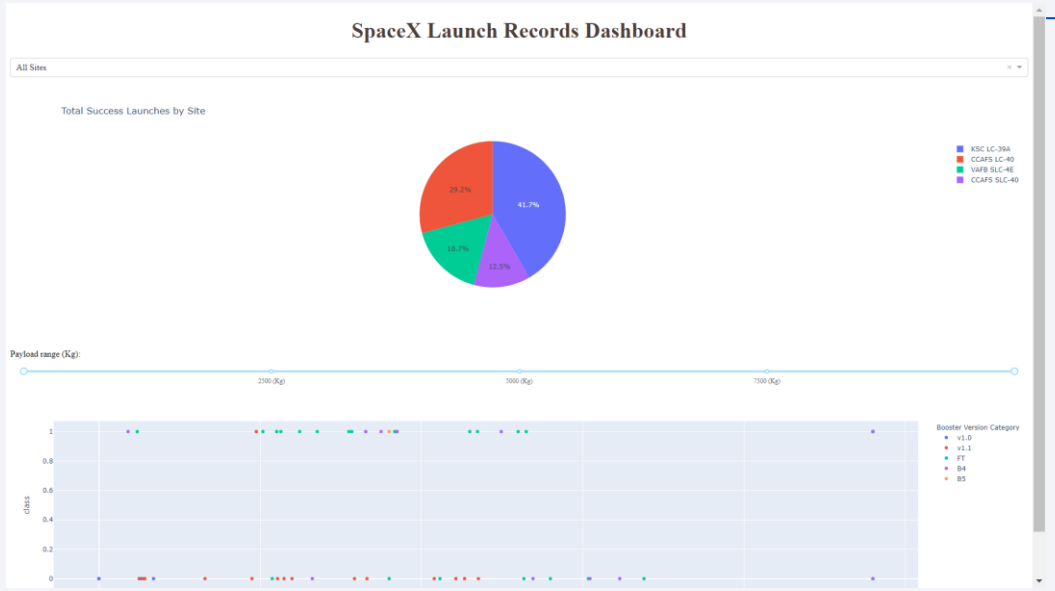
https://github.com/bsensix/IBM_Data_Science_Professional_Certification/blob/main/Module_3/spacex_dash_app.py

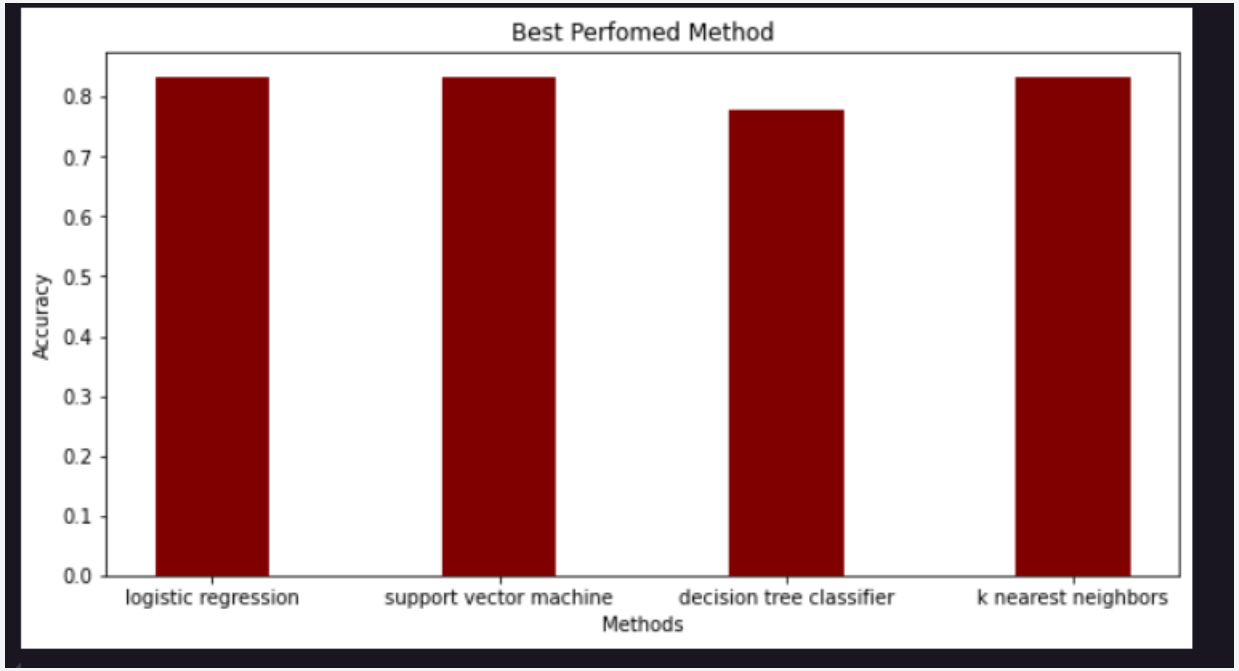
Predictive Analysis (Classification)

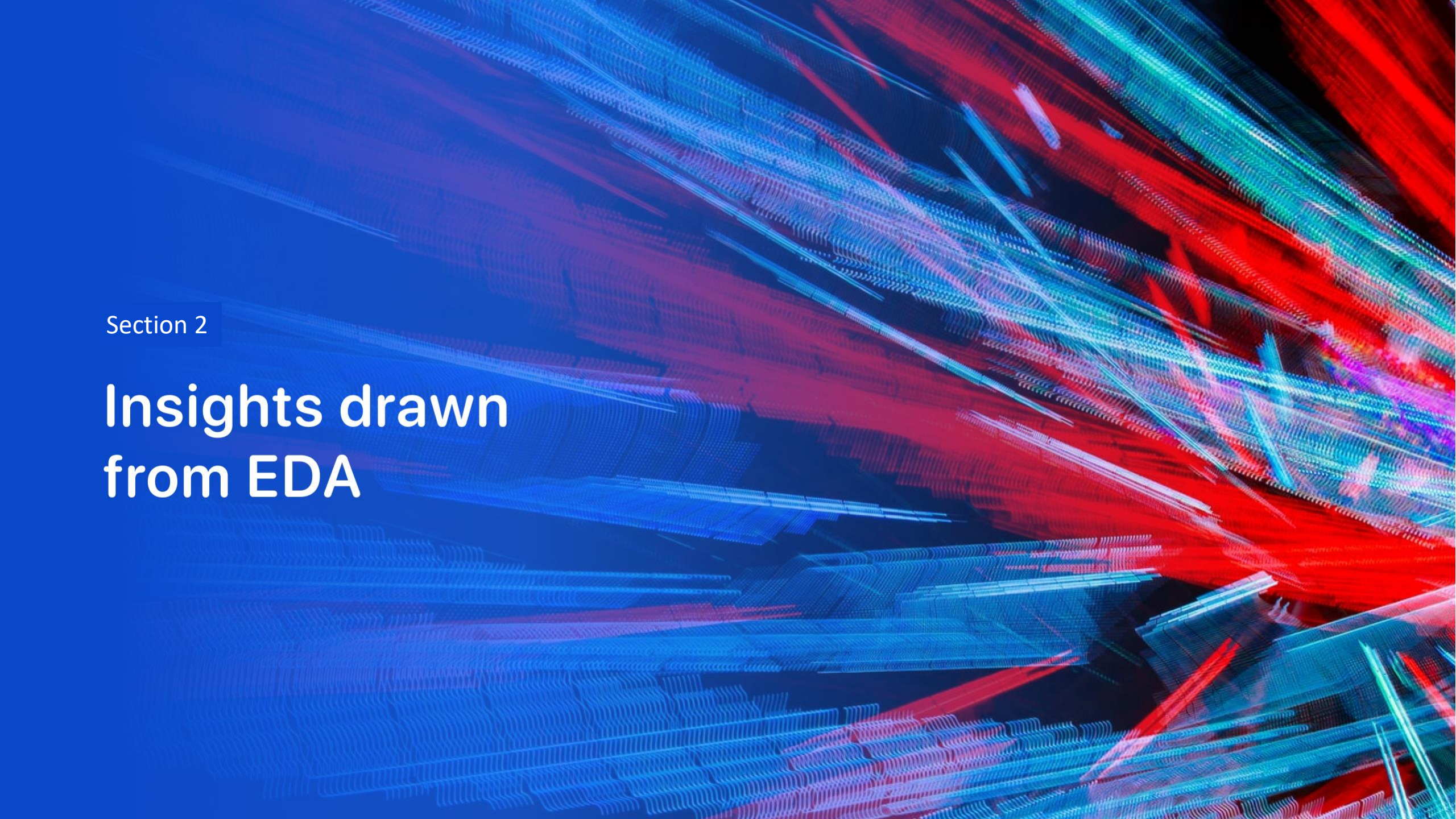
- Add the GitHub URL: https://github.com/bsensix/IBM_Data_Science_Professional_Certification/blob/main/Module_4/machine_learning_prediction.ipynb



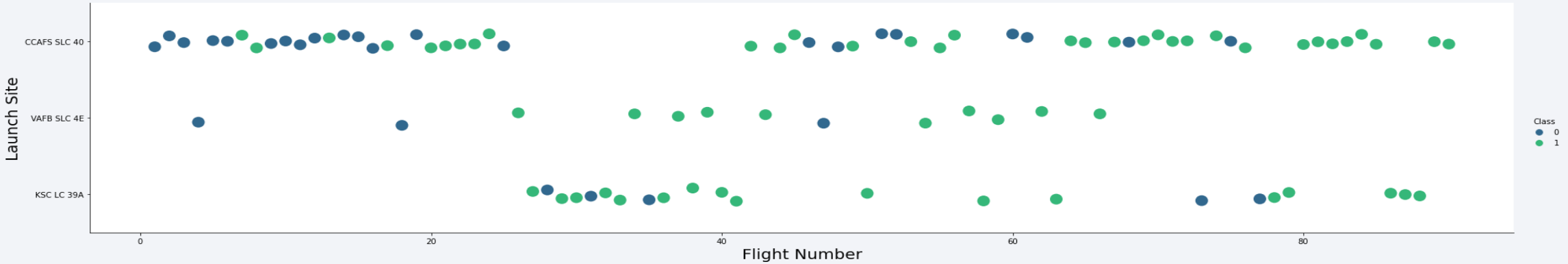
Results


SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
29.2%
41.7%
16.7%
12.5%
Payload range (Kg):
2500 (Kg)
5000 (Kg)
7500 (Kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5
class


Best Perfomed Method
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
logistic regression
support vector machine
decision tree classifier
k nearest neighbors
Methods

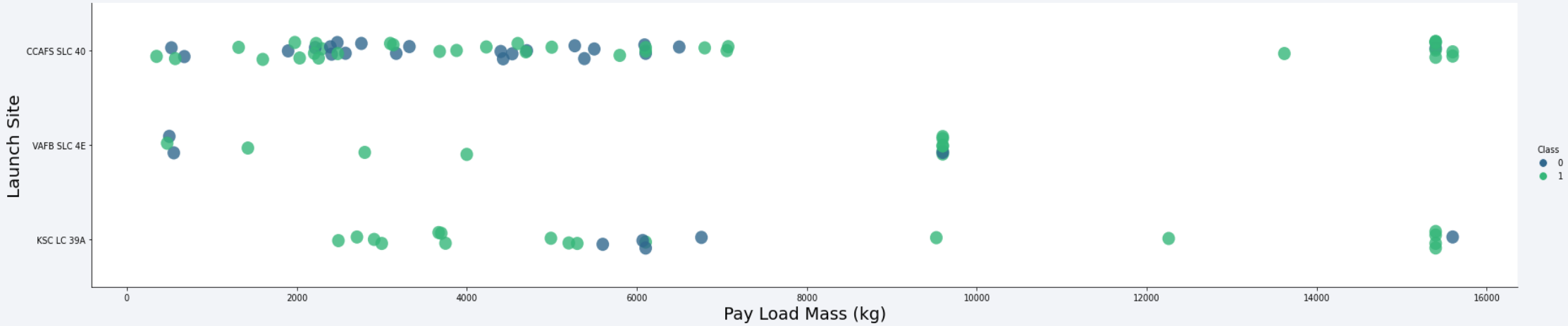

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



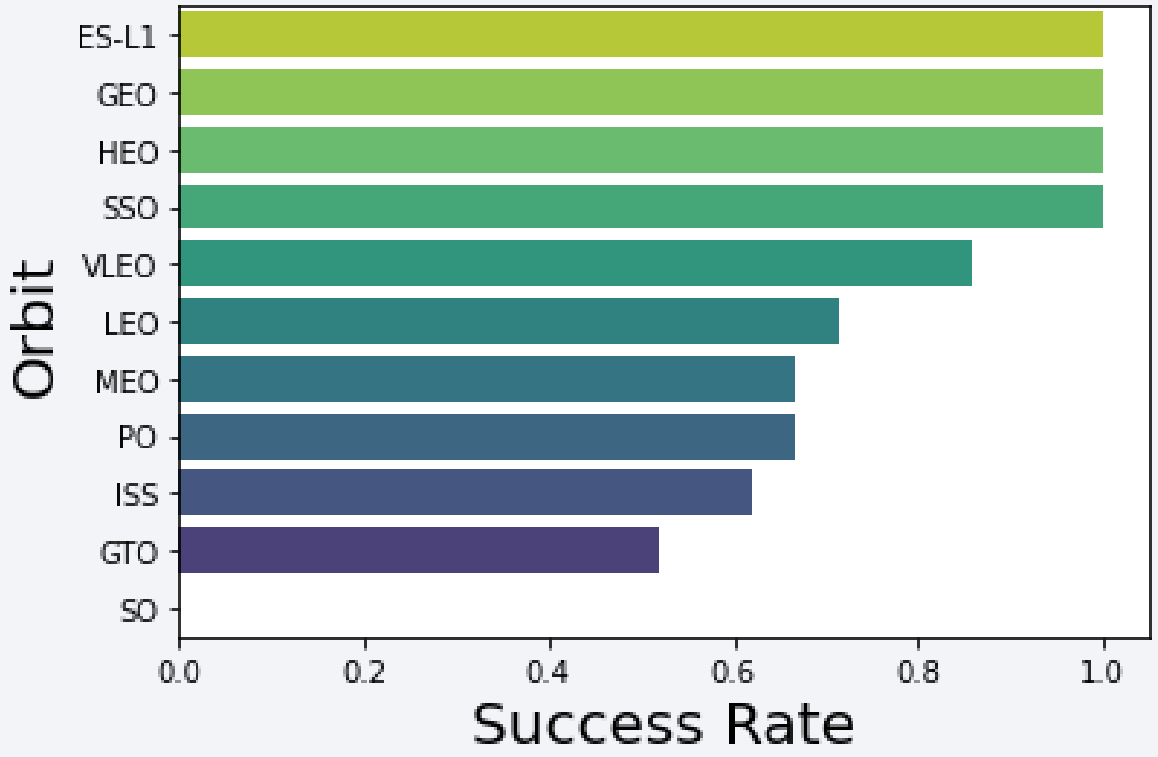
The graph indicates a rise in success rates over time (as indicated by Flight Number). There seems to have been a significant breakthrough around the 20th flight that notably boosted success rates. CCAFS appears to be the primary launch site, given its higher volume.

Payload vs. Launch Site



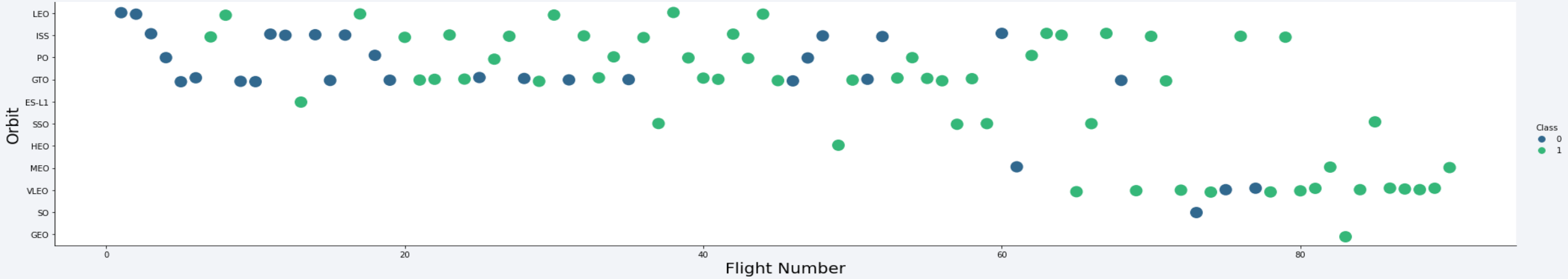
The payload mass mostly ranges between 0-6000 kg. Different launch sites also appear to handle varying payload masses.

Success Rate vs. Orbit Type



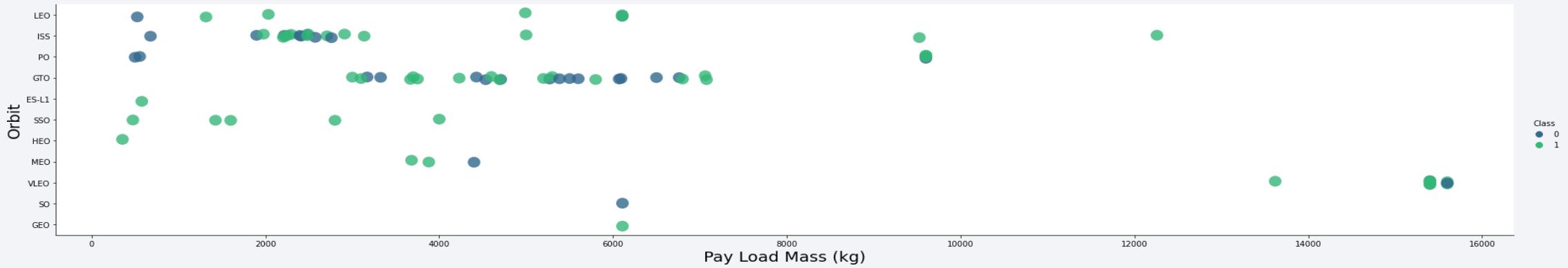
ES-L1 (1), GEO (1), HEO (1) have a 100% success rate (sample sizes in parentheses). SSO (5) also boasts a 100% success rate. VLEO (14) shows a decent success rate with several attempts. SO (1) has a 0% success rate. GTO (27) exhibits around a 50% success rate and represents the largest sample size.

Flight Number vs. Orbit Type



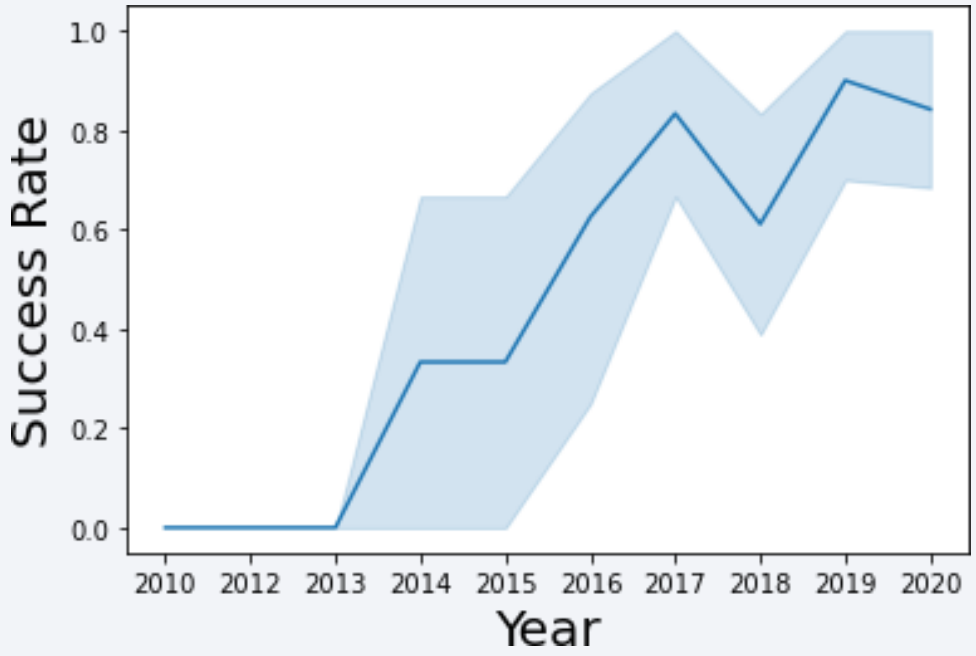
Launch orbit preferences shifted over flight numbers, and launch outcomes seem to correlate with these preferences. SpaceX initially focused on LEO orbits, which saw moderate success. They then transitioned back to VLEO in recent launches. SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

Payload vs. Orbit Type



Payload mass appears to correlate with orbit type. LEO and SSO orbits tend to have relatively low payload masses. In contrast, VLEO, which is another successful orbit, typically involves payload mass values at the higher end of the range.

Launch Success Yearly Trend



Success rates have generally increased since 2013, with a slight dip in 2018. In recent years, success rates have stabilized around 80%.

All Launch Site Names

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f2
        Done.
```

Out[4]:

launch_site
CCAFS LC-40
CCAFS SLC-40
CCAFSSLC-40
KSC LC-39A
VAFB SLC-4E

Query unique launch site names from database.

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors.

CCAFS LC-40 was the previous name.

Likely only 3 unique launch_site values:

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

Launch Site Names Begin with 'CCA'

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

sum_payload_mass_kg
45596

This query sums the total payload mass in kg where NASA was the customer.

CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8
Done.
```

avg_payload_mass_kg
2928

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

This query returns the first successful ground pad landing date.

First ground pad landing wasn't until the end of 2015.

Successful landings in general appear starting 2014

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.database
Done.
```

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

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

mission_outcome	no_outcome
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

This query provides a count of each mission outcome. SpaceX seems to achieve its mission outcomes nearly 99% of the time. This suggests that most landing failures are deliberate. Interestingly, one launch has an unclear payload status, and unfortunately, one ended in flight failure.

Boosters Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

This query identifies the booster versions that carried the highest payload mass of 15600 kg. These booster versions are very similar and all belong to the F9 B5 B10xx.x variety. This likely suggests that payload mass correlates with the specific booster version used.

2015 Launch Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
Done.
```

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

In 2015, there were two occurrences where the first stage failed to land on a drone ship.

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

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

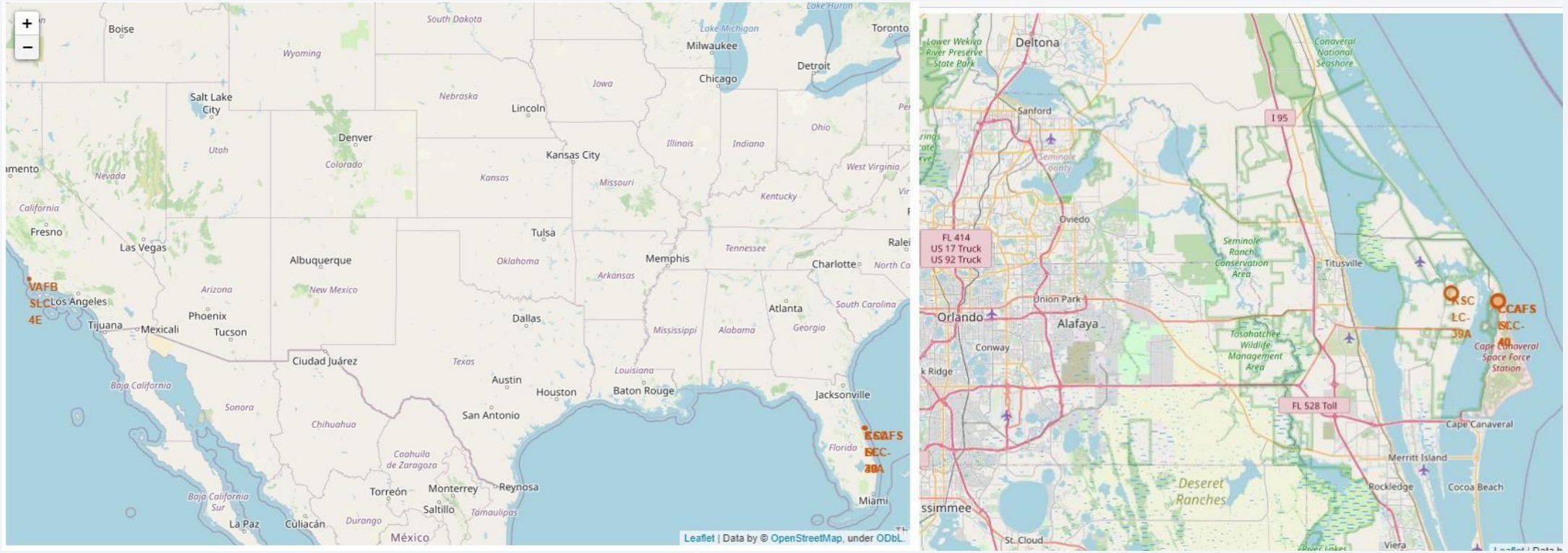
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

Between June 4, 2010, and March 20, 2017, inclusive, there were a total of 8 successful landings. These successful landings include both drone ship and ground pad landings.

Section 3

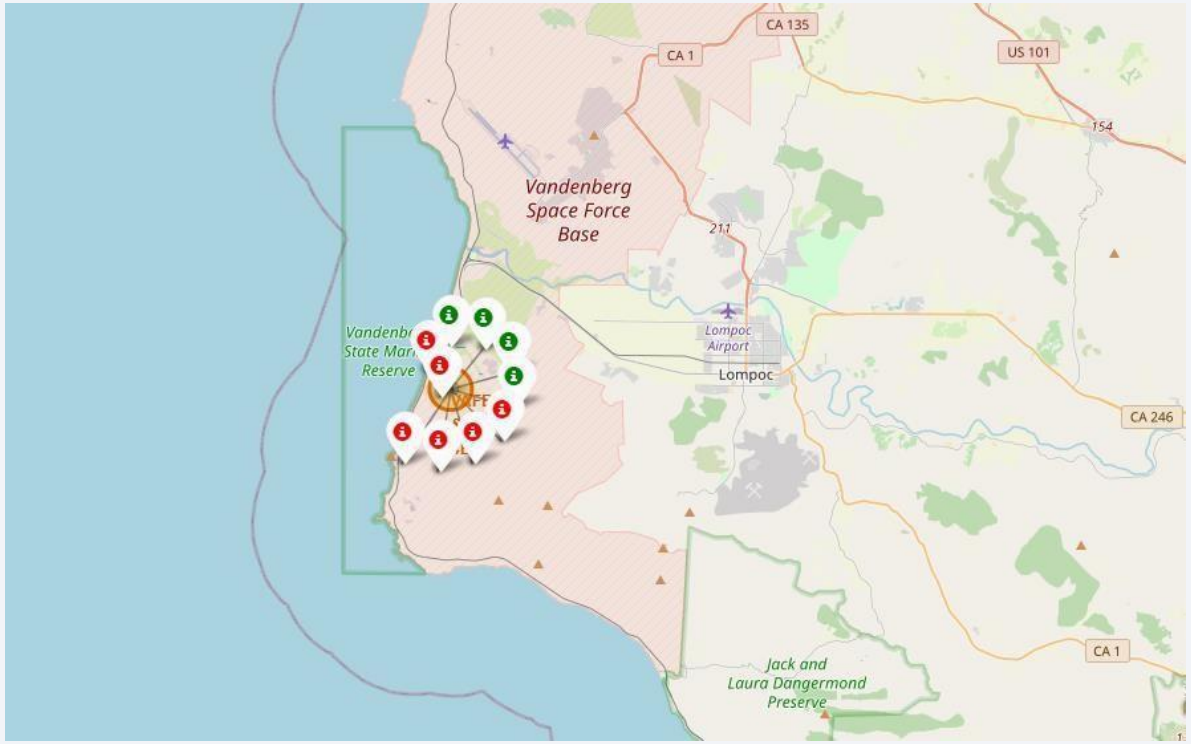
Launch Sites Proximities Analysis

<Folium Map Screenshot 1>



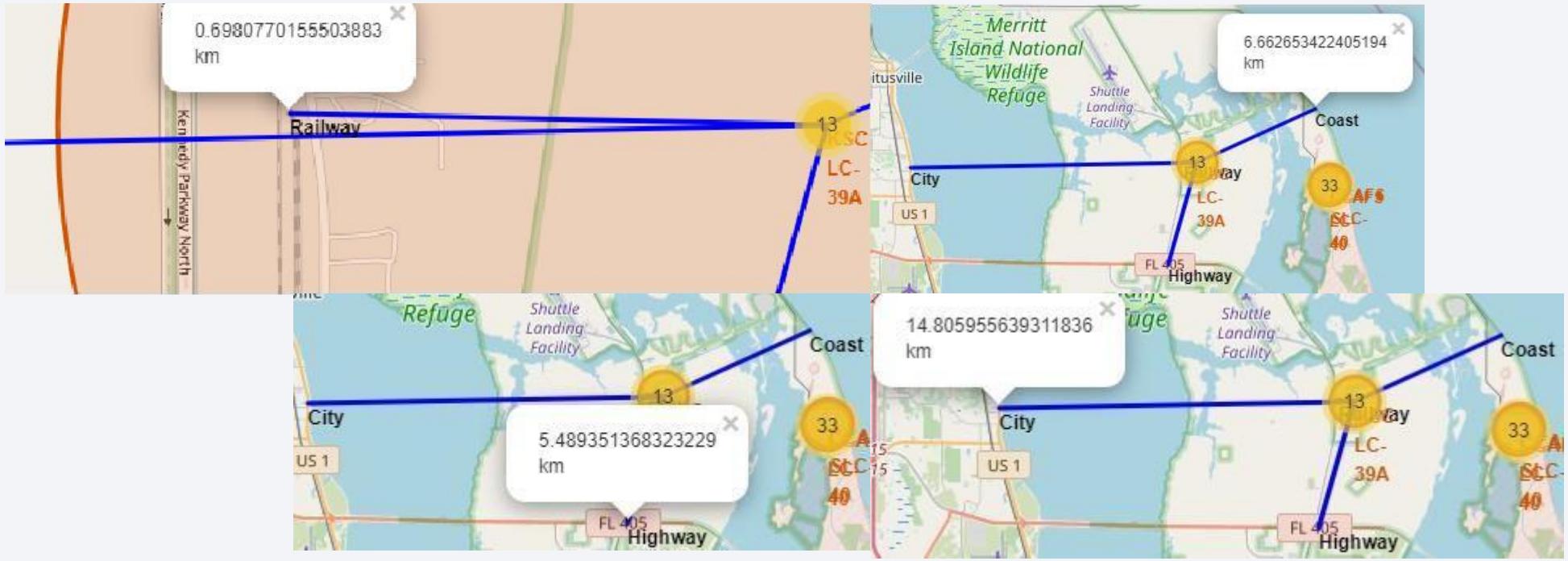
The map on the left displays all launch sites relative to the US map. The map on the right specifically highlights the two Florida launch sites due to their proximity to each other. All launch sites are situated near the ocean.

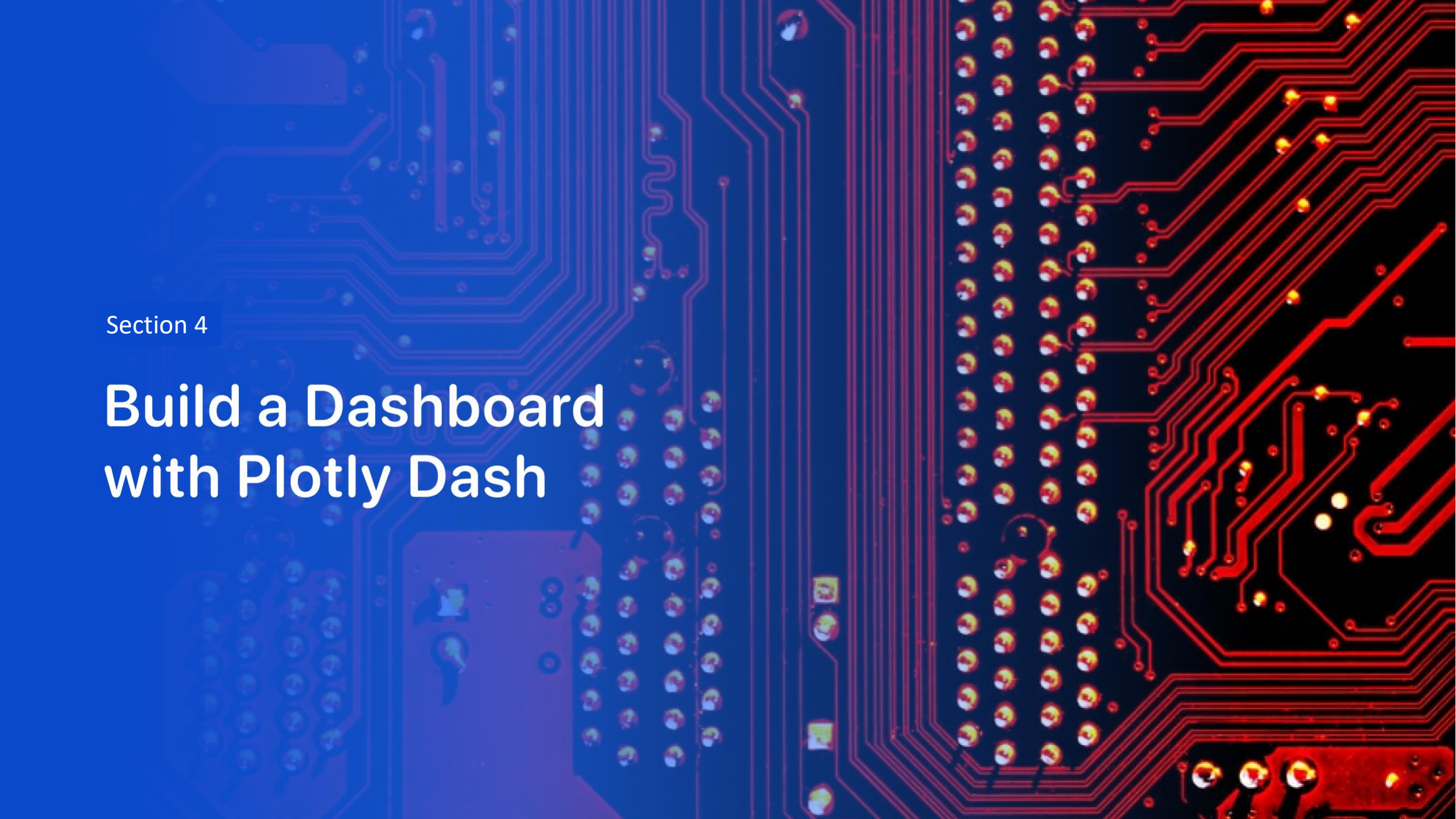
<Folium Map Screenshot 2>



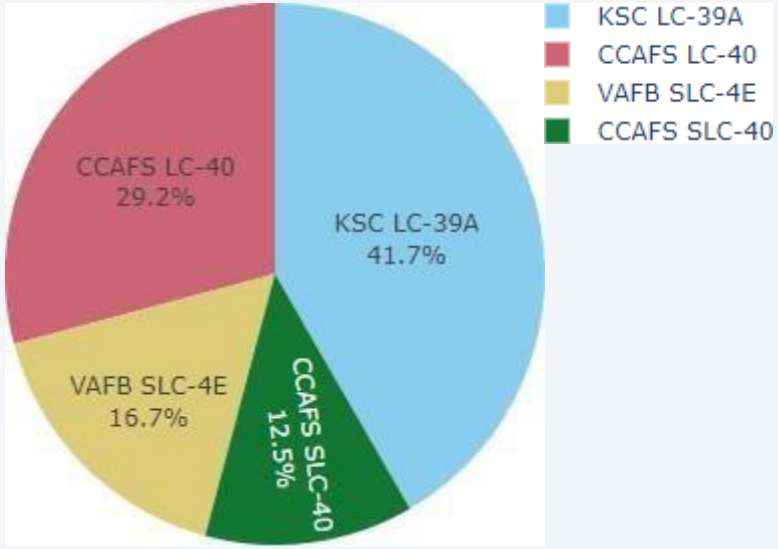
Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon).

<Folium Map Screenshot 3>



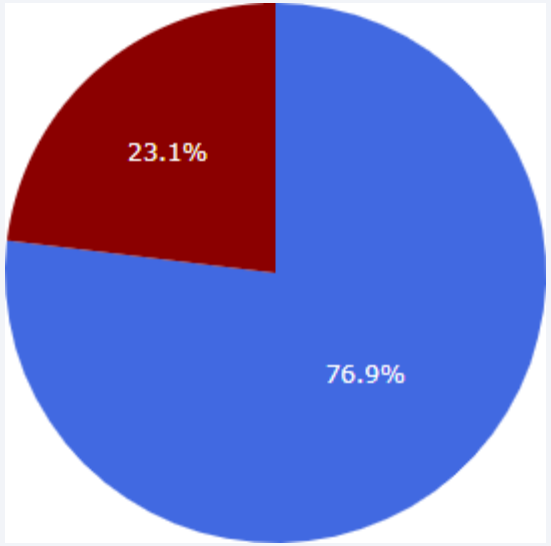

Section 4
Build a Dashboard with Plotly Dash

<Dashboard Screenshot 1>



This distribution shows the successful landings across all launch sites. CCAFS LC-40, which is the old name of CCAFS SLC-40, has the same number of successful landings as KSC. However, a majority of these successful landings occurred before the name change. VAFB has the smallest share of successful landings, possibly due to a smaller sample size and increased difficulty in launching from the West Coast.

<Dashboard Screenshot 2>



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

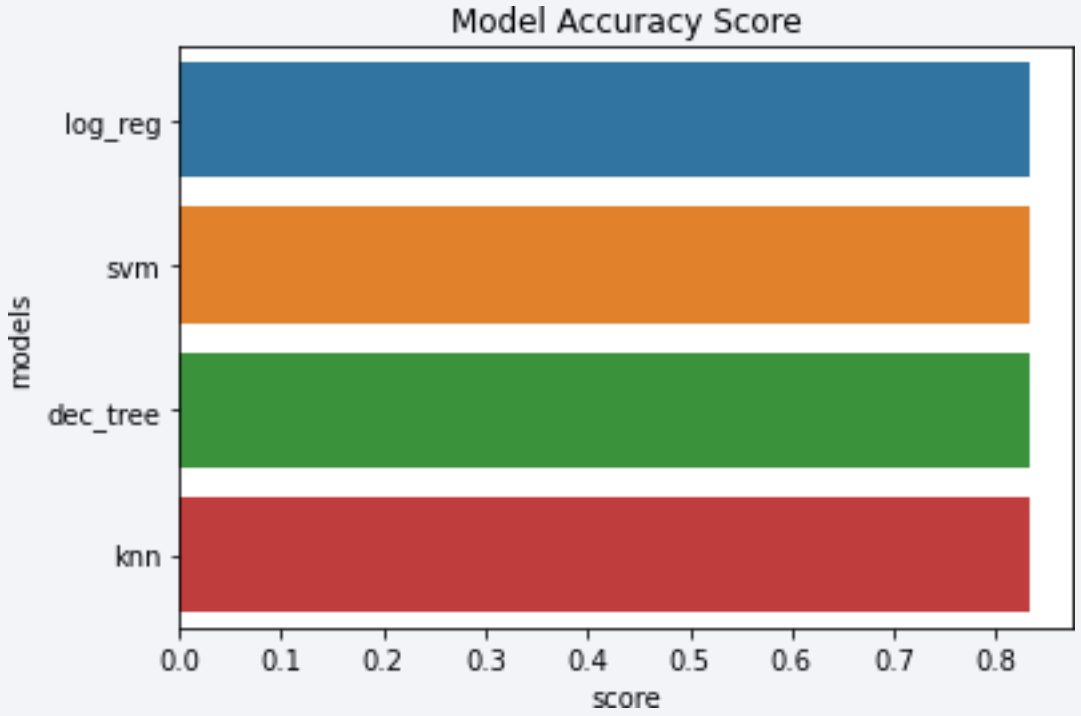
<Dashboard Screenshot 3>

The Plotly dashboard includes a payload range selector, but it's currently set from 0 to 10000 kg instead of the maximum payload of 15600 kg. The "Class" indicator uses 1 for successful landings and 0 for failures. The scatter plot also incorporates the booster version category for color and the number of launches for point size. Interestingly, within the range of 0-6000 kg, there are two instances of failed landings with payloads of zero kilograms.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



All models achieved nearly the same accuracy on the test set, with an average of 83.33%. It's important to note that the test size is small, consisting of only 18 samples. This small sample size can lead to significant variance in accuracy results, as observed in repeated runs of the Decision Tree Classifier model. To determine the best model more reliably, additional data would likely be necessary.

Confusion Matrix

Since all models performed equally on the test set, the confusion matrix is identical across all models. The models correctly predicted 12 successful landings when the true label was a successful landing. They also predicted 3 unsuccessful landings when the true label was an unsuccessful landing. However, the models incorrectly predicted 3 successful landings when the true label was an unsuccessful landing (false positives), indicating that the models tend to overpredict successful landings.

Conclusions

Our task was to develop a machine learning model for Space Y, aiming to predict when Stage 1 will successfully land to potentially save ~$100 million USD. We utilized data from a public SpaceX API and web scraping the SpaceX Wikipedia page. After creating data labels and storing them in a DB2 SQL database, we developed a dashboard for visualization purposes. The machine learning model we created achieved an accuracy of 83%.

With this model, Allon Mask of SpaceY can predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch. This prediction can help determine whether the launch should proceed or not.

To further enhance the model's performance, it would be beneficial to collect more data. This additional data can help refine the machine learning model selection and improve overall accuracy.

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!