IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alexandre Leblanc-Richard

2023-04-16

Outline

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

Executive Summary

- Space Y wishes to determine the best parameters for a successful landing
- To determine these parameters, a series of analysis was performed using data available from the Space X Rest API and the Space X Wikipedia page
    - The data was preprocessed to format it and fill missing values
    - Analysis were performed with visuals using Pandas dataframes in Python and also using the sqlalchemy library with a db2 database
    - A plotly dashboard was built to visualize the data interactively
    - Folium maps were created to visualize the locations of the different launch sites around the USA, and to determine the proximities
- With this analysis, we have been able to determine two factors that appear to impact the success of a landing:
    - Launching from the Kennedy Space Center
    - Using one these orbits in conjunction with a payload mass below 6,000 kg: ES-L1, GEO, HEO and SSO
    - The VLEO orbit also has a high success rate with payloads greater than 13,000 kg
- We then tested four different machine learning models to determine which would best predict the outcome of a landing
- All four models tested yielded the same accuracy on the testing data, at 83%
- The best model that was identified is the decision tree classifier, with an accuracy of 89% on the training data

Introduction

- Project background and context
  - Many companies wish to make commercial space flights available and affordable
  - One of the most successful => Space X
    - Can reuse the expensive first stage, cutting costs by more than half ($165M vs $62M per launch)
    - Depending on mission parameters, the first stage can land successfully, crash or be sacrificed
  - Space Y wants to compete with Space X using Space X's results to determine the best course of action
- Problems you want to find answers to
  - Using a machine learning model, can we predict the outcome of a flight given certain parameters
  - Can we identify optimal parameters for a successful launch

Section 1

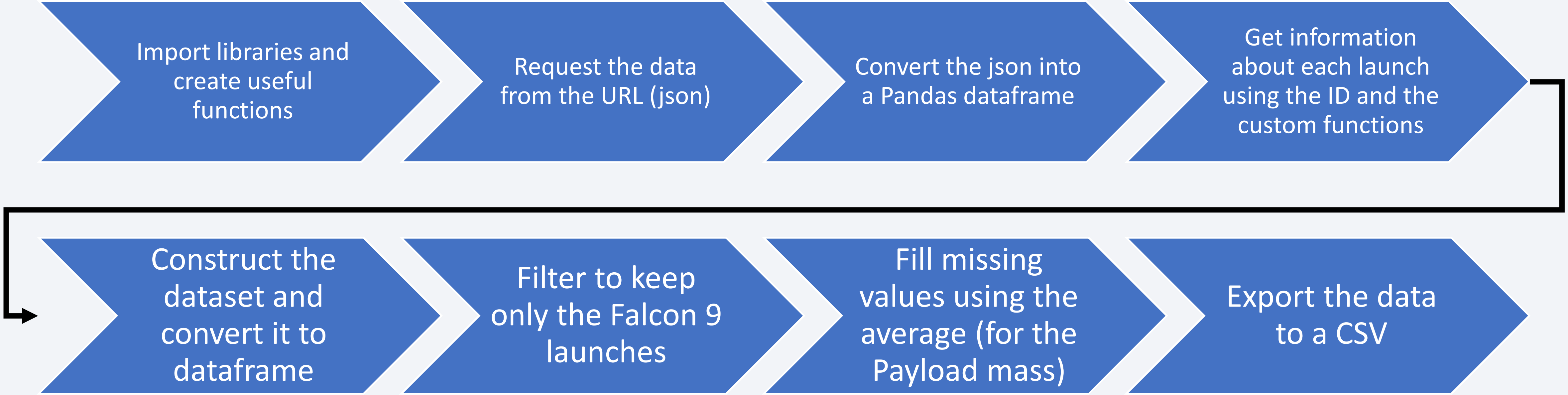
Methodology

Methodology

Executive Summary

- Data collection methodology:
    - Space X Rest API
    - Wikipedia page on Space X Falcon 9 Heavy Launches
- Perform data wrangling
    - Filtering and filling missing values
    - Add landing class to the data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - GridSearch CV
    - Bar chart to rank models

Data Collection – SpaceX API



- Extraction was performed using the SpaceX API to get information about the Falcon 9 launches
- We used the Request library in Python to query the API and construct the dataset
- Link: https://github.com/AlexandreLR/CapstoneProjet/blob/579947c7ac412eb9b24f404360d0a7b0b0605b3f/Data%20collection.ipynb

Data Collection – Web scraping

1. Import libraries and create useful functions
2. Use the Request library to get the content of the page and create a BeautifulSoup object
3. Extract the column headers/variable names from the table header
4. Parse the launch tables to get the different information and store the contents in a dictionnary
5. Convert the dictionary into a dataframe and export as a CSV

- Extraction was performed from the Wikipedia page
- We used the Request and BeautifulSoup libraries in Python to extract the data from the HTML tables in the page
- Link: https://github.com/AlexandreLR/CapstoneProjet/blob/579947c7ac412eb9b24f404360d0a7b0b0605b3f/Data%20collection%20-%20Web%20scraping.ipynb

Data Wrangling

Import libraries the previously extracted data → Explore some statistics on the data: number of launch by site, by orbit and by mission outcome → For the mission outcome, create a set of bad outcomes → Using the bad outcome set, add a landing_class variable (1 or 0) → Export data into CSV format

- Data contains different types of mission outcomes
- Need to convert these string outcomes into labels we can use for training (1 for success, 0 for failure)
- Link: https://github.com/AlexandreLR/CapstoneProjet/blob/579947c7ac412eb9b24f404360d0a7b0b0605b3f/Data%20wrangling.ipynb

EDA with Data Visualization

- Plotted charts:
    1. Scatter plots to show the relation and present any potential correlation between the variables
        1. Payload mass vs flight number
        2. Launch site vs flight number
        3. Launch site vs payload mass
        4. Orbit vs flight number
        5. Orbit vs payload mass
    2. Bar chart to compare success rate by orbit
    3. Line chart to show the evolution of the success rate over the years
- Link: https://github.com/AlexandreLR/CapstoneProjet/blob/6e0c996875a17000371bf832d191715f7519ee6f/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- List of queries performed
  - Names of the different launch sites
  - Display 5 records for launch sites starting with CCA (using LIKE and wildcard “%”)
  - Total payload mass carried by boosters launched by NASA
  - Average payload mass for the Falcon 9 booster version 1.1
  - List the boosters with success for landing on drone ships with a payload mass between 4,000 and 6,000 kg
  - List the total number of mission outcomes by outcome type (success or failure)
  - List of booster version that carried the maximum payload mass
  - List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015
  - Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order

- Link: https://github.com/AlexandreLR/CapstoneProjet/blob/3b8e3b9d60e068326eda4ae4ace226cfcc348bd2/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

1. Add map markers for launch sites to view the location of all the launch sites
2. Mark success and failures by launch site by adding a MarkerCluster object. This allows to quickly view which sites have the most success
3. Calculate the launch site distance to its proximities using a map marker and an icon to show the distance on the map along with a line. This will allow us to determine where an ideal launch site should be positionned

- Link: https://github.com/AlexandreLR/CapstoneProjet/blob/dcd76cf82edb4ed1b3a2a1515b3cbaada81b9fe5/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

Build a Dashboard with Plotly Dash

- Added a pie chart to show the success rate by launch site to determine which launch site has the most success
- Added a scatter plot to show the relation between then payload mass, the success rate and the booster version. This allows us to determine which version of the booster is best suited for each payload mass category.
- Both visuals can be filtered by launch site for more details
- Link: https://github.com/AlexandreLR/CapstoneProjet/blob/9411639f541e755bfac61b073234865ca83f625f/spacex_dash_app.py

Predictive Analysis (Classification)

1. Import libraries and create a function to plot the confusion matrix
2. Import the data and perform a standard scaling
3. Split the data into training and testing sets (80-20)
4. Test 4 models using GridSearch: Logistic regression, SVM, decision tree, KNN
5. Plot a confusion matrix for each model
6. Calculate the accuracy score for the test data
7. Plot the accuracy scores for all models, for both the training and testing data
8. Determine the best model given the accuracy score

- Each model was tested using GridSearchCV and a list of different parameters depending on the model used
- Link: https://github.com/AlexandreLR/CapstoneProjet/blob/0299a11386718c1d99b3de5d9f6e6b440cfdb007/Machine%20learning%20prediction%20lab.ipynb

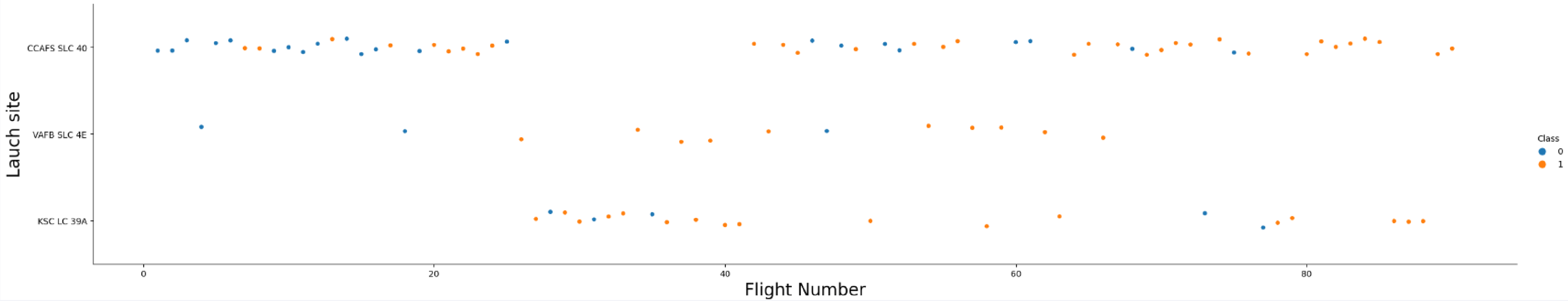
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

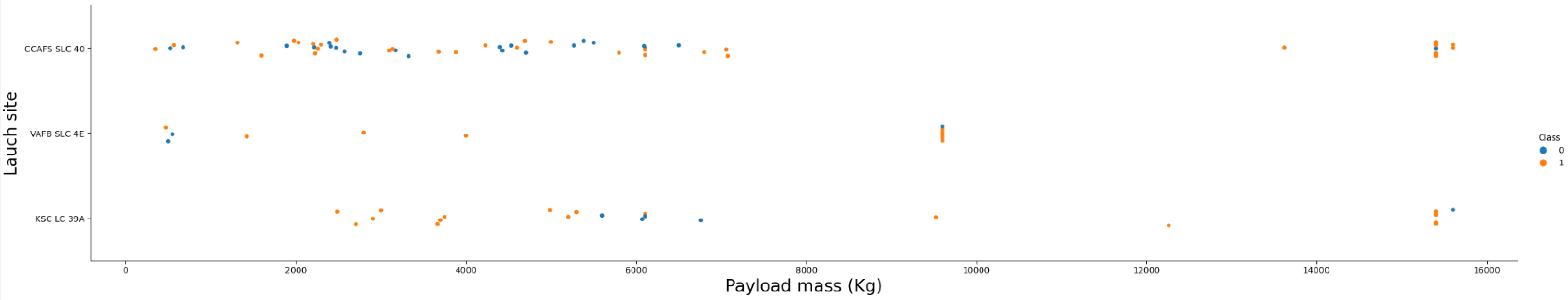
Insights drawn from EDA

Flight Number vs. Launch Site



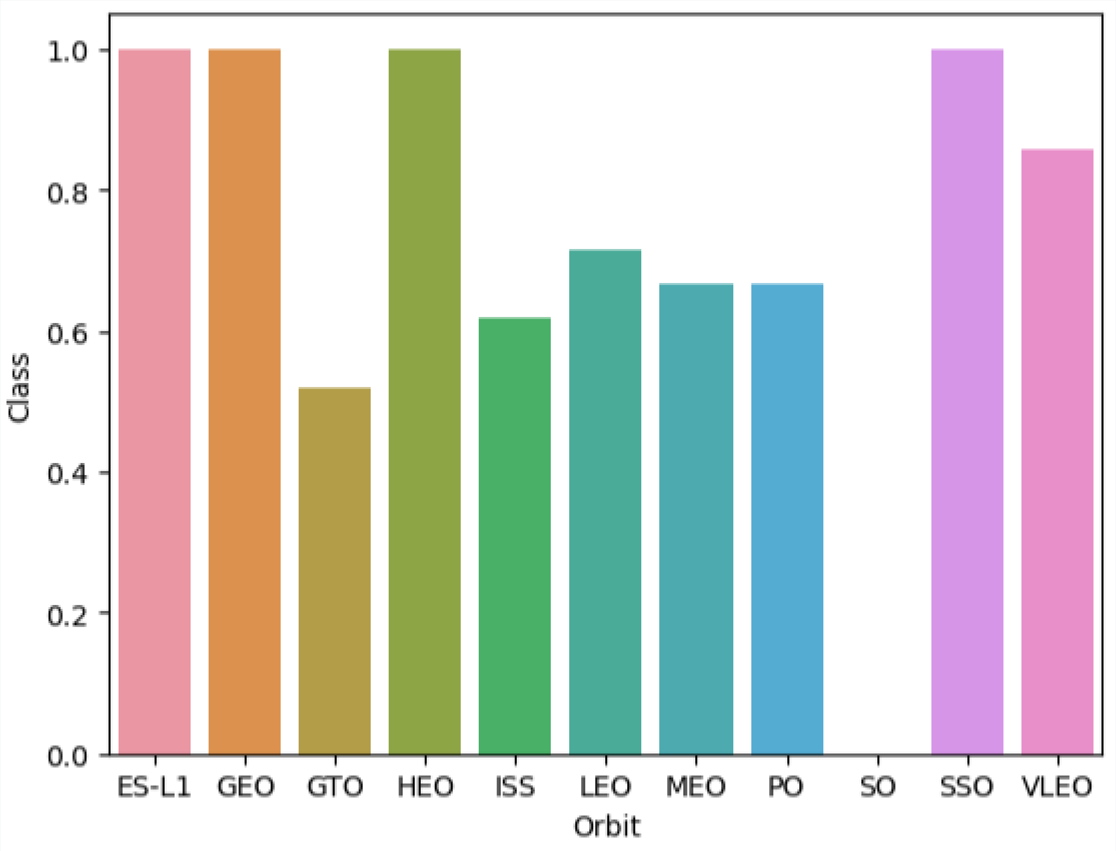
- Early flights for CCAFS-SLC-40 => Low success rate
- After 20th flight => Success rate increased
- No failed flights after the 80th flight

Payload vs. Launch Site



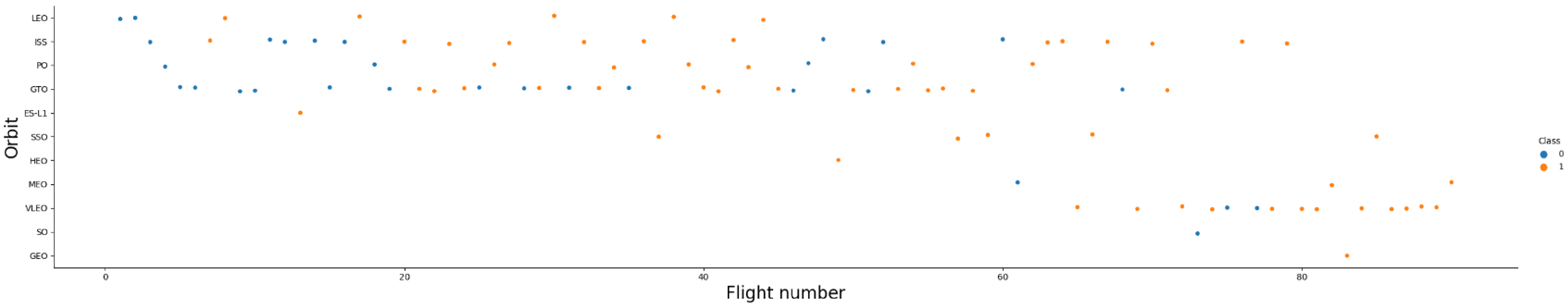
- The site VAFB SLC 4E has not launched payloads with as mass greater than 10,000 kg
- The success rate appears to be higher for payload masses above 10,000 kg
- KSC-LC-39A had only one failures for payload masses below 6,000 kg

Success Rate vs. Orbit Type



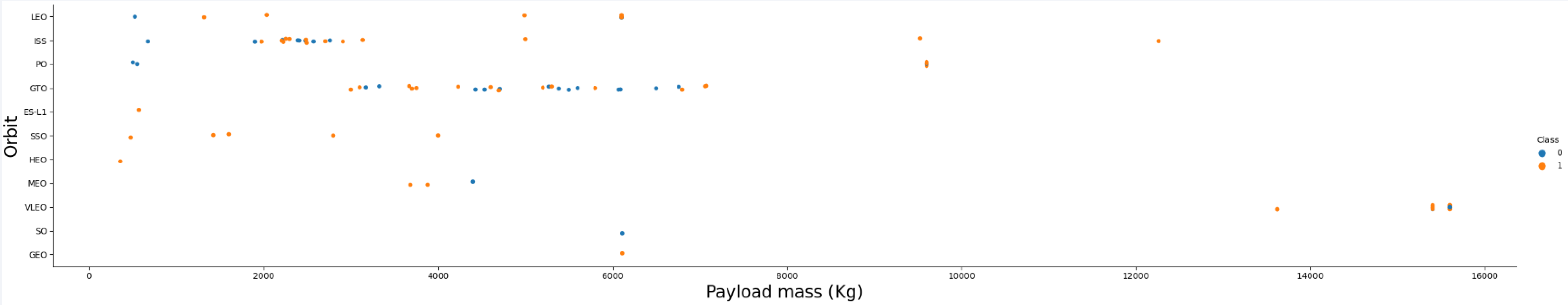
- ES-L1, GEO, HEO and SSO have the highest success rates (100%)
- Need to confirm in multiple launches for these orbits (see next slide)
- GTO has the lowest success rate

Flight Number vs. Orbit Type



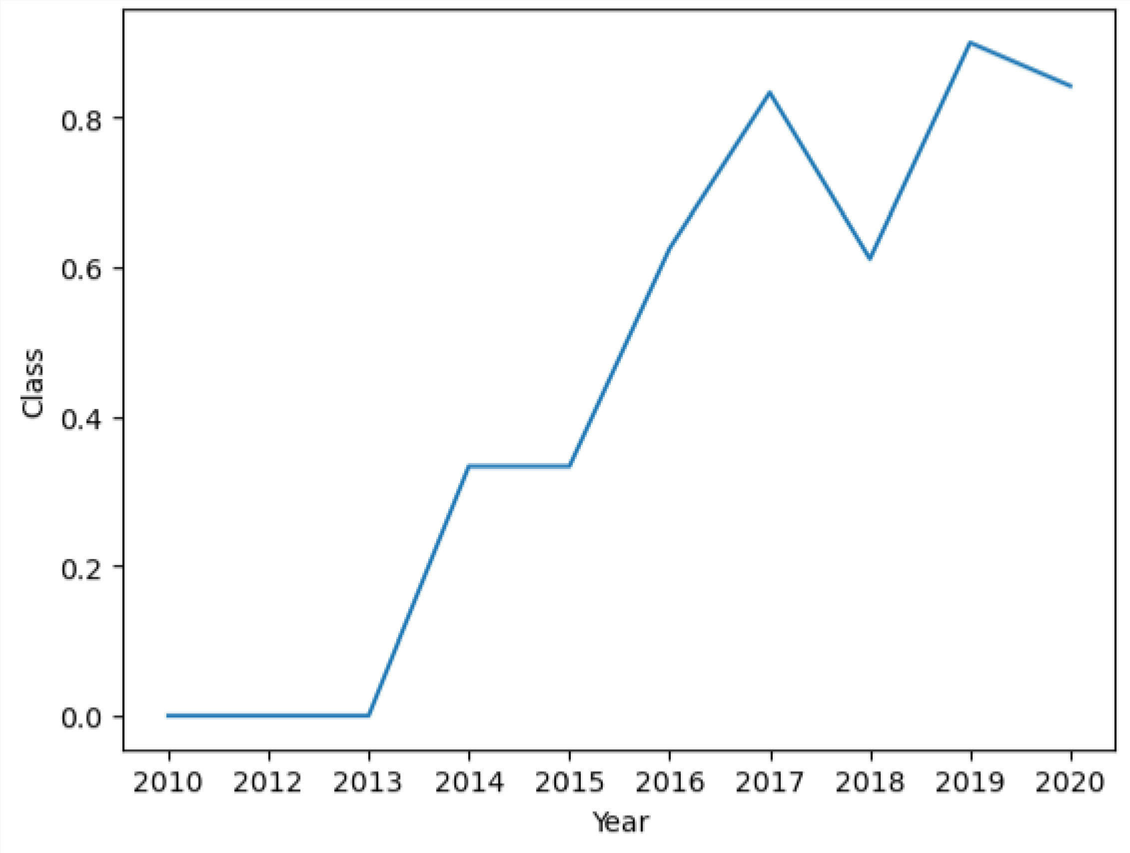
- LEO and PO orbits' success rate appear to be related to the flight number, while for other orbits the correlation is not as pronounced
- Some orbits only have very few flights (such as HEO, SO or GEO), which explains some high or low success rates

Payload vs. Orbit Type



- aaa

Launch Success Yearly Trend



- The success rate since 2013 increased each year except in 2018, which saw a decrease

All Launch Site Names

```
%%sql
SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL
```

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

]:

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

- There are a total of 4 launch sites used by Space X:
  - 2 in Cape Canaveral Launch Complex, Florida
    - CCAFS LC-40
    - CCAFS SCL-40
  - 1 at the Kennedy Space Center, Florida
    - KSC LC-39A
  - 1 at the Vadenberg Space Launch Complex (California)
    - VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
%%sql
SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
2010-06-04 00:00:00	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08 00:00:00	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 00:00:00	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 00:00:00	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01 00:00:00	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

- Out of first five launches in Cape Canaveral, all missions were successful
- In 3 out 5 mission, no landing was attempted, while in the other 2 the landing was a failure

Total Payload Mass

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE CUSTOMER="NASA (CRS)"
```

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

SUM(PAYLOAD_MASS__KG_)
45596

- The total payload mass carried for NASA (CRS) was 45,596 kg

Average Payload Mass by F9 v1.1

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE BOOSTER_VERSION = "F9 v1.1"
```

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

]:

AVG(PAYLOAD_MASS__KG_)
2928.4

- The average payload mass for the Falcon 9 booster version 1.1 was 2,928 kg

First Successful Ground Landing Date

```
%%sql
SELECT MIN(DATE) FROM SPACEXTBL WHERE "LANDING _OUTCOME"="Success (ground pad)"
```

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

]:

MIN(DATE)
2015-12-22 00:00:00

- The first successful landing on a ground pad was in December of 2015, on the 22nd

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE "LANDING _OUTCOME"="Success (drone ship)" AND PAYLOAD_MASS__KG_>4000 AND PAYLOAD_MASS__KG_<6000
```

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

]:

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

- For payload masses between 4,000 and 6,000 kg, 4 booster version had successful landings on a drone ship

Total Number of Successful and Failure Mission Outcomes

```
%%sql
SELECT MISSION_OUTCOME,COUNT(*) AS TotalNumber FROM SPACEXTBL GROUP BY MISSION_OUTCOME
```

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

3]:

Mission_Outcome	TotalNumber
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- Out of all the mission outcomes (not landing outcome), 99 were success, while only 2 appear to have had issues (1 failure in flight and 1 were the status of the payload is unclear)

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT Booster_Version,PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

- The maximum payload mass carried by a Space X booster was 15,600 kg
- Here is a list

2015 Launch Records

```
%%sql
SELECT SUBSTR(Date,6,2) AS Month,"Landing _Outcome",Booster_Version,Launch_Site FROM SPACEXTBL WHERE SUBSTR(Date,1,4)='2015' AND "Landing _Outcome"='Failure (drone ship)'
```

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

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- In 2015, 2 missions had failed landings on a drone ship (one in January and one in April)
- The 2 were launched from Cape Canaveral Launch Complex 40
- There were 2 booster version: F9 v1.1 B1012 and F9 v1.1 B1015

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

```
%%sql
SELECT LandingOutcome,Cnt FROM (
    SELECT "Landing _Outcome" AS LandingOutcome,COUNT(*) AS Cnt FROM SPACEXTBL WHERE Date>='2010-06-04' AND Date<='2017-03-20' GROUP BY "Landing _Outcome"
) ORDER BY Cnt DESC
```

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

5]:

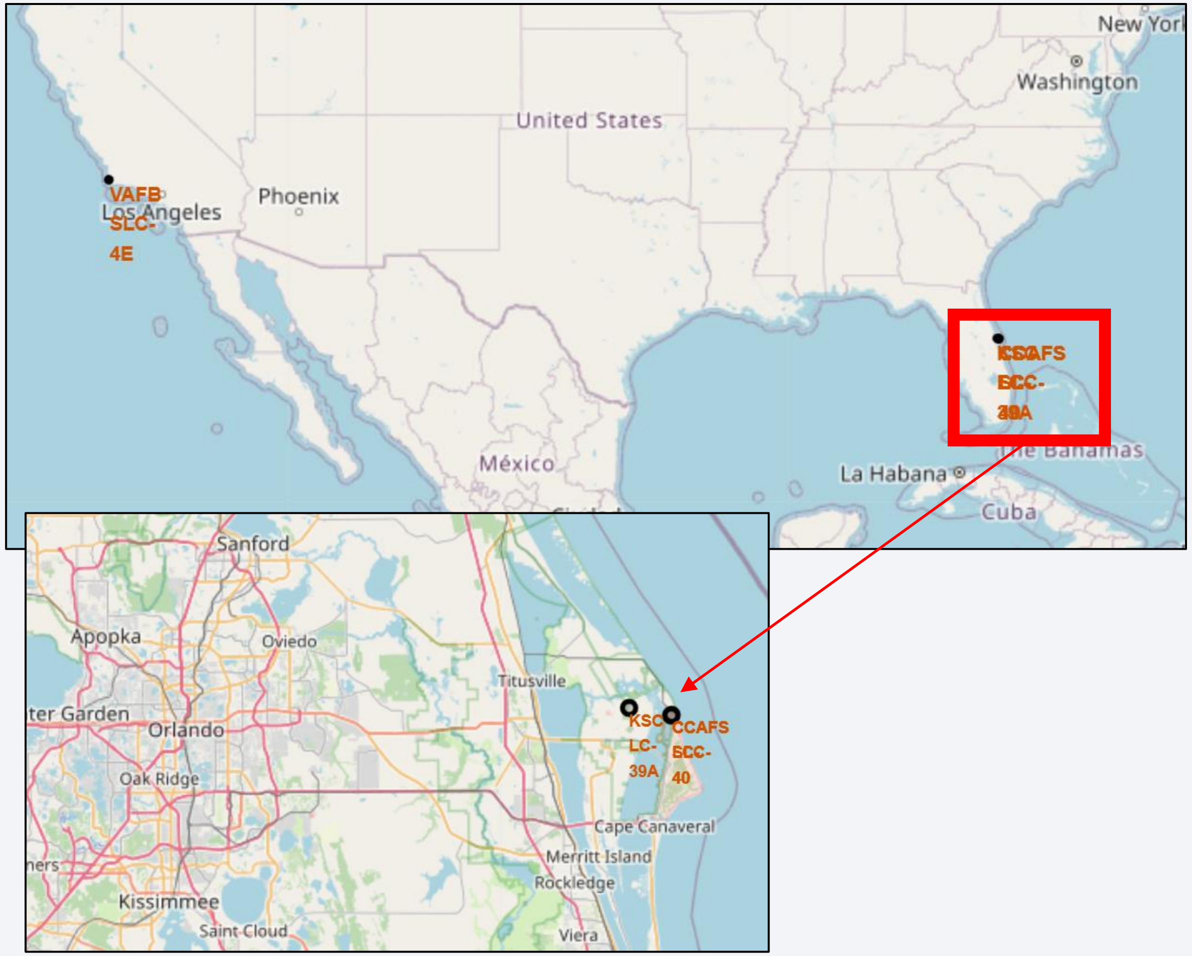
LandingOutcome	Cnt
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- Between June 4, 2010 and March 20, 2017, there were a total of 31 flights
  - 10 had no attempted landings and 7 were failures
  - 50% of the drone ship landings failed

Section 3

Launch Sites Proximities Analysis

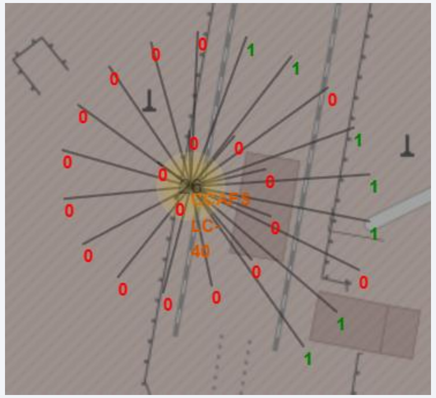
Launch site locations



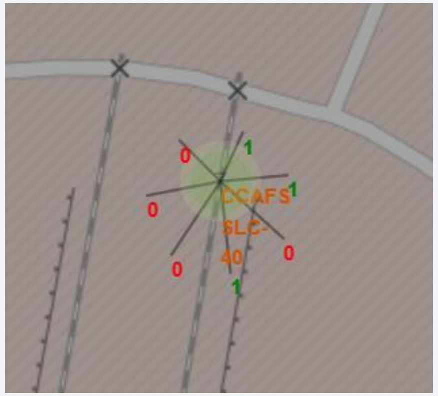
- Launch sites are primarily located on coasts
- One launch site on the west coast and 3 on the east coast
- Launch sites are in proximity of the equator

Launch outcomes by launch site

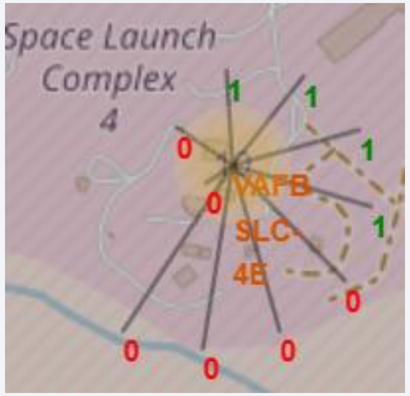
CCAFS-LC-40



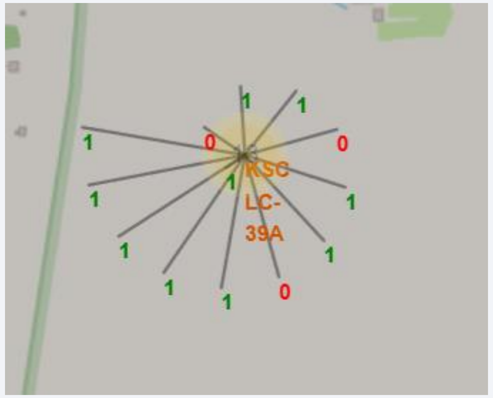
CCAFS-SLC-40



KSC-LC-39A

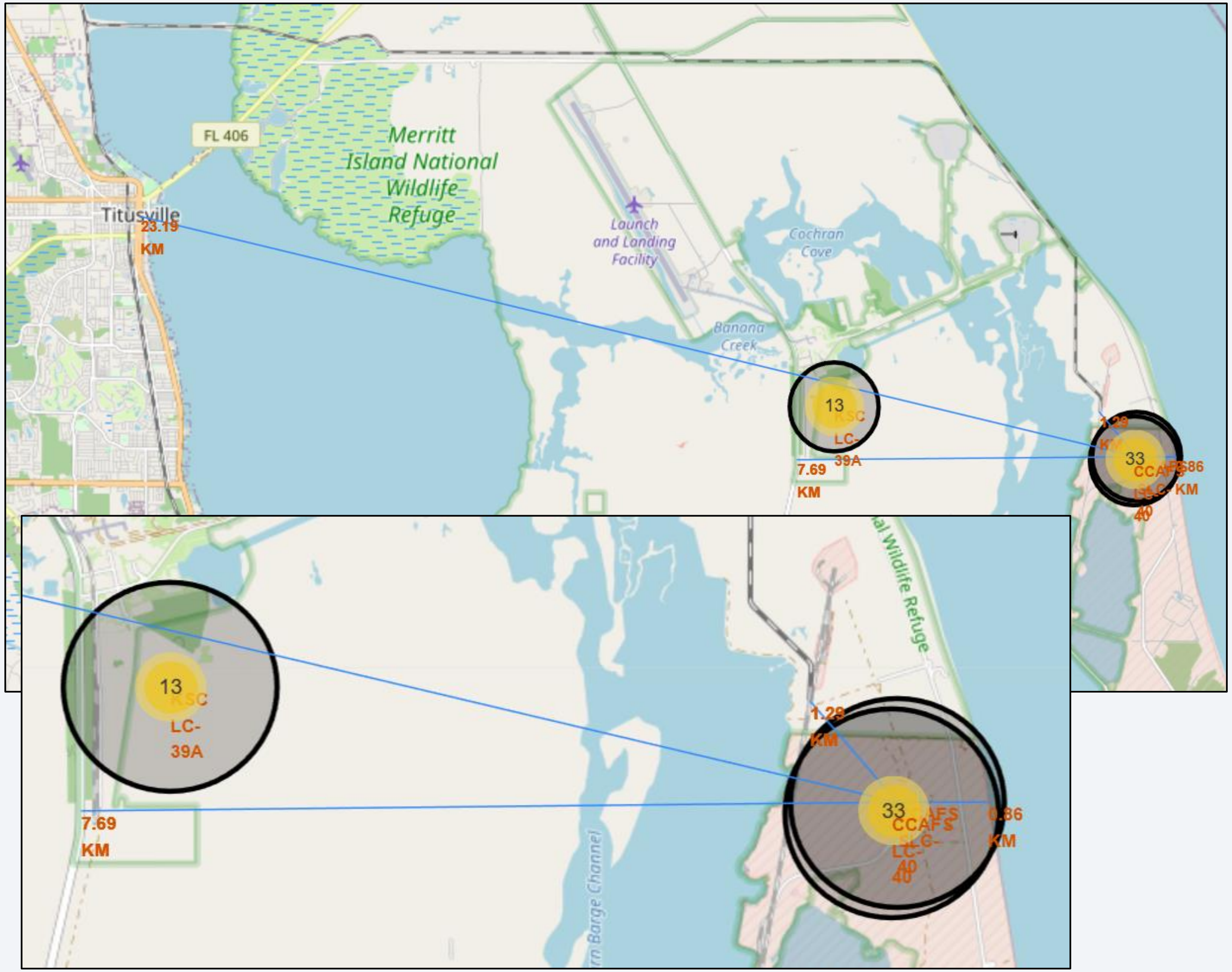


KSC-LC-39A



- Added labels for each of the launch by launch site to show the mission outcomes
- KSC-LC-39A has the most successful launch of all site
- CCAFS-LC-40 has the most failed launches

Launch site proximities

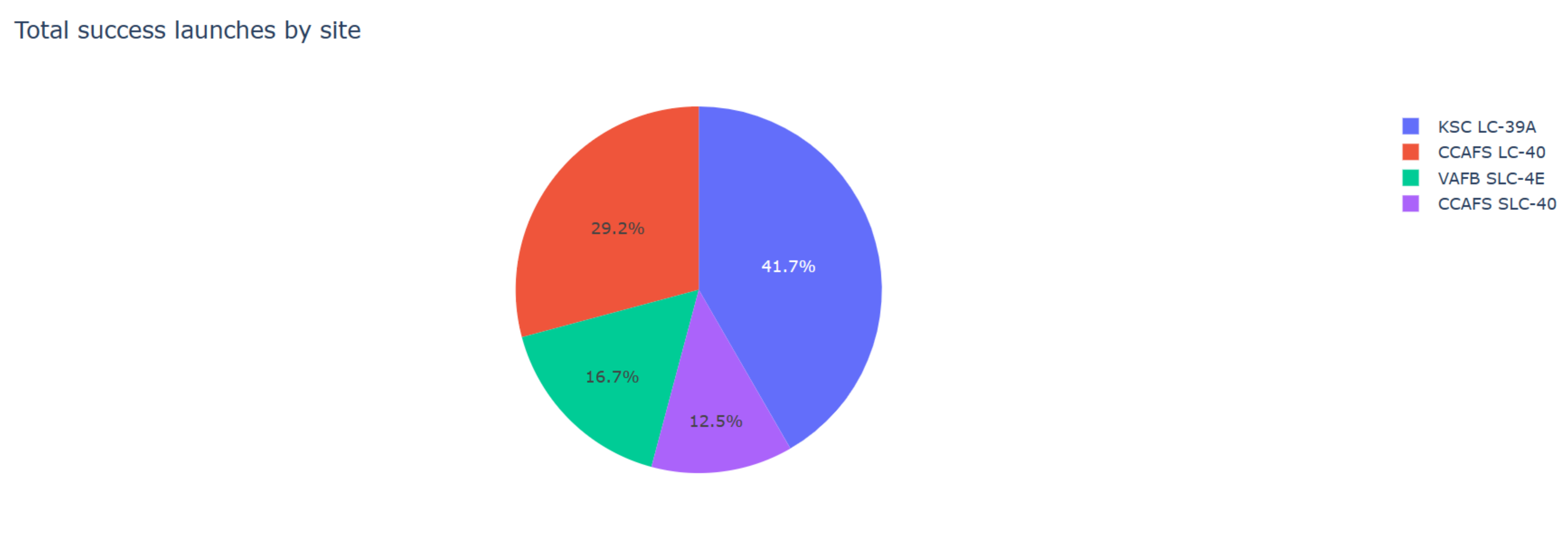


- The launch sites are located close to the ocean, with a 0.86 km distance for CCAFS-SLC-40
- Also close to railways (1.23 km)
- Situated relatively far from cities (23.19 km from Titusville)
- Hard to conclude whether or not launch sites are close to highways (7.69km for CCAFS-SLC-40, but much closer to KSC-LC-39A)

Section 4

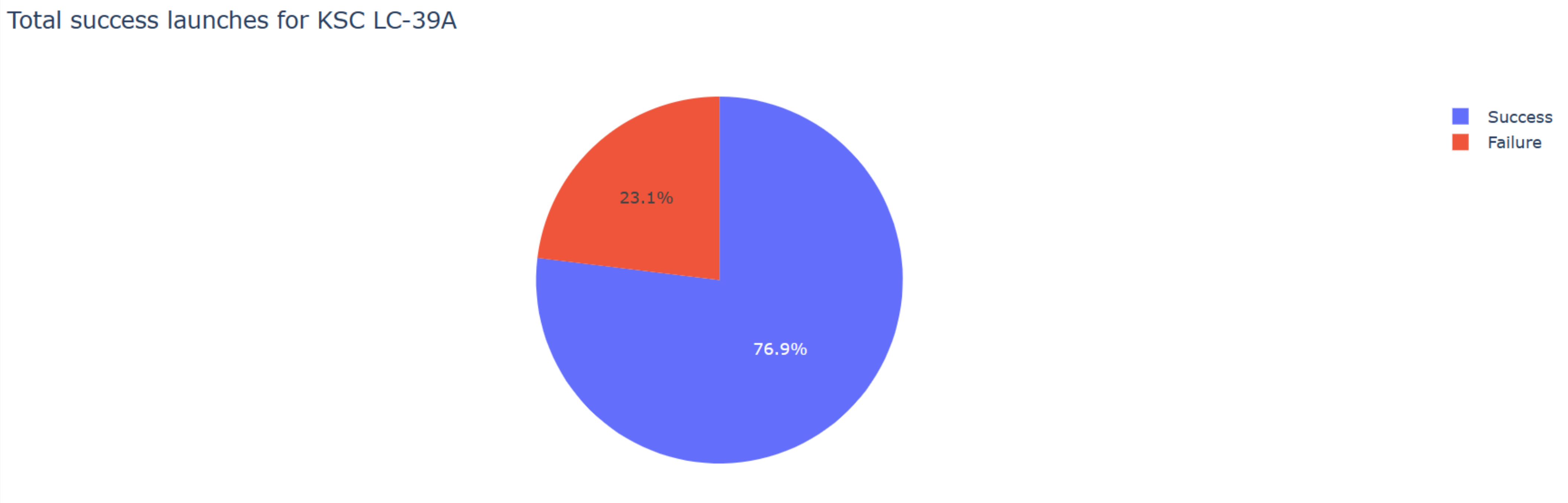
Build a Dashboard with Plotly Dash

Success by launch sites



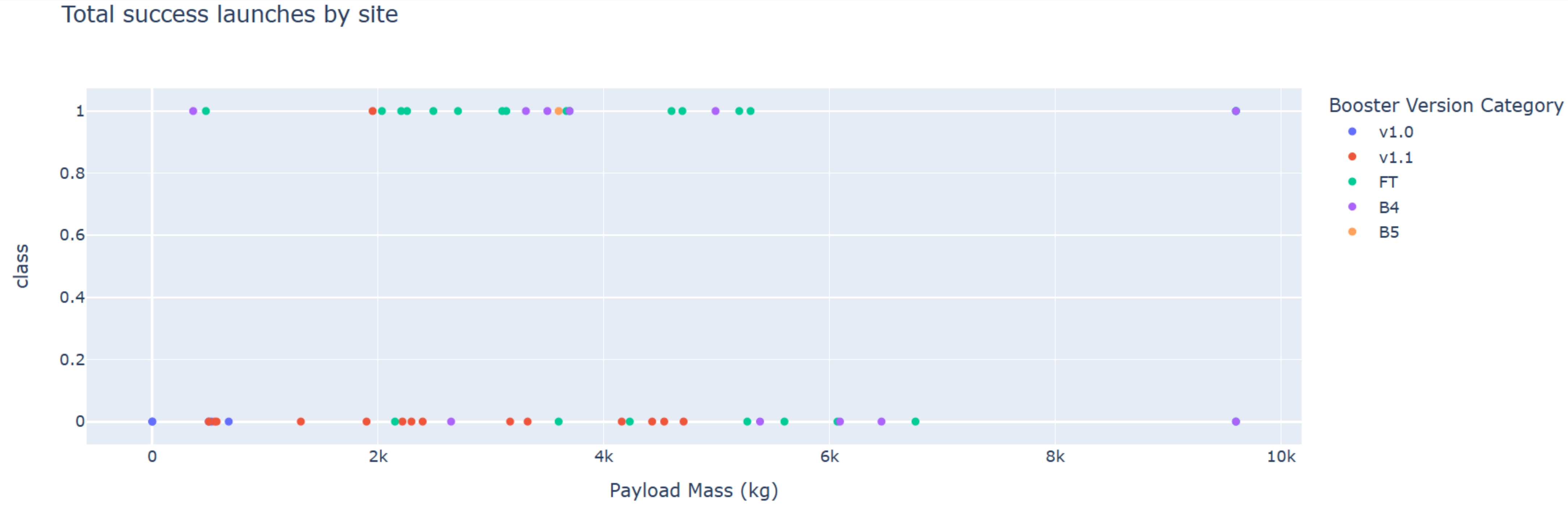
- KSC LC-39A has the highest success rate of all sites, which confirms what we saw in the Folium map before

Success rate for KSC LC-39A



- KSC LC-39A has over 75% success launches

Success rate by payload mass and booster version



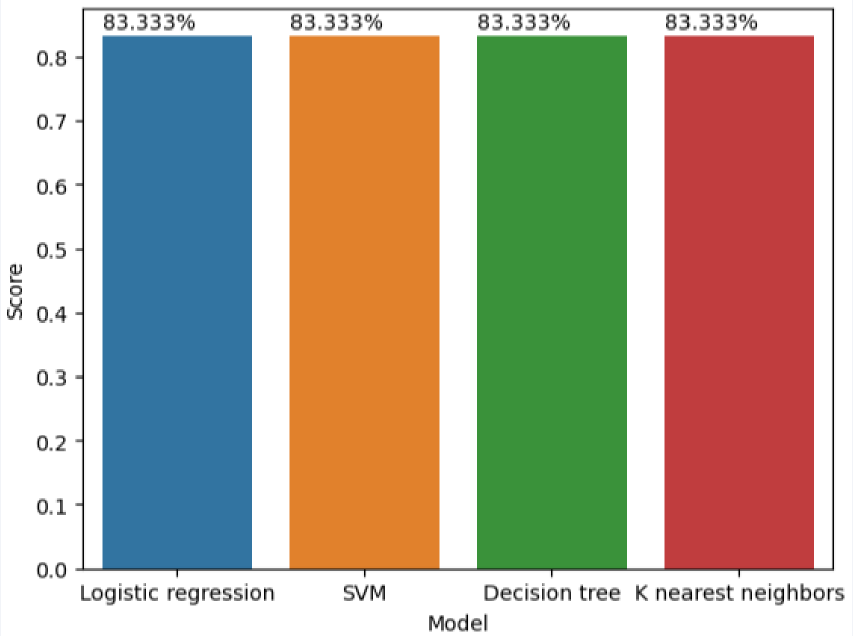
- FT booster version has a high success rate for payloads under 6,000 kg
- The v1.1 booster version has a low success rate regardless of the payload mass

Section 5

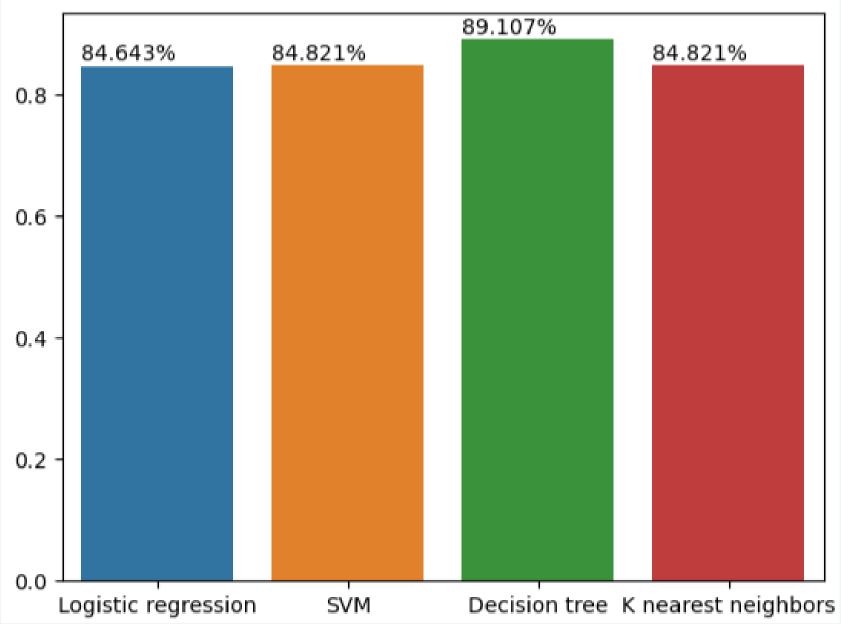
Predictive Analysis (Classification)

Classification Accuracy

Accuracy score – Test data

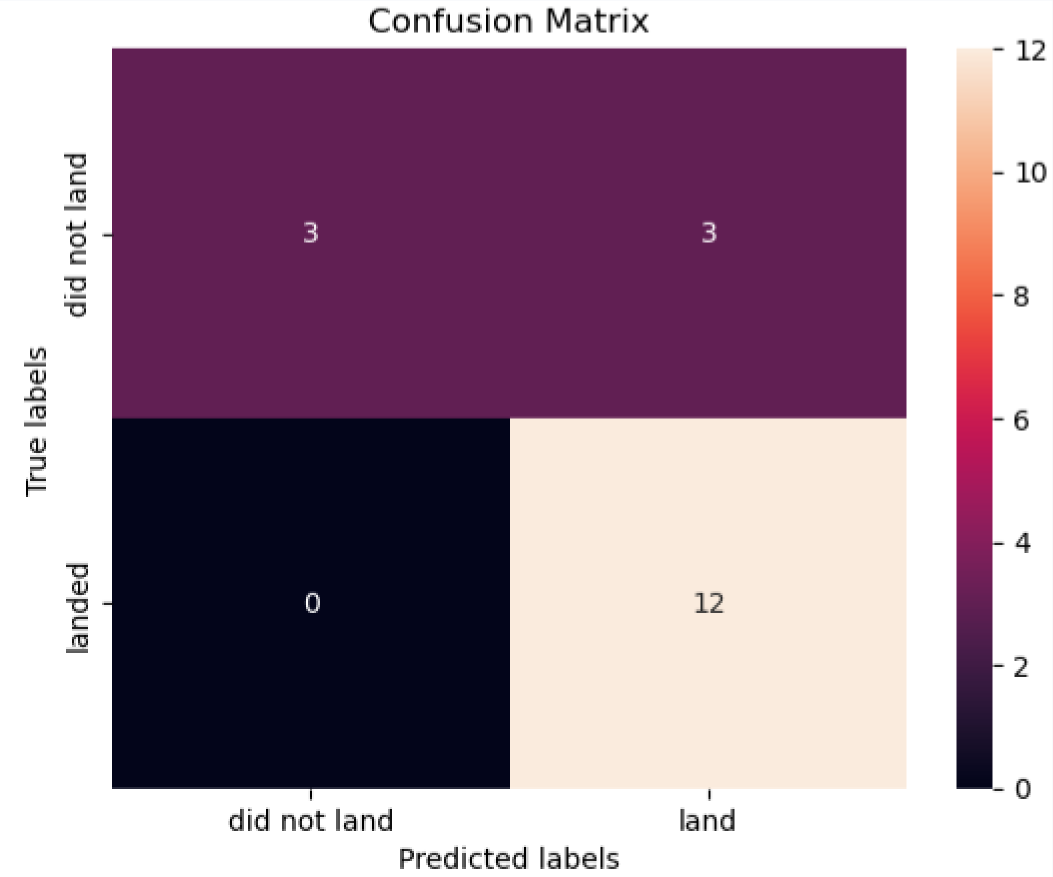


Accuracy score – Training data



- Since all models had the same accuracy score on the test data, we had to turn to the accuracy of the training data to get the best model
- Best model => Decision tree classifier

Confusion Matrix



- The decision tree model is very good to predict a landing when one really occurred
- It is however unable to accurately predict when the landing did not happen
  - 50% of the time it predicted wrong

Conclusions

- The questions we wanted answered were:
  - Can we predict the outcome of a launch using a machine learning model?
  - What are the optimal parameters for a successful landing?
- Given all the data we had at our disposal, here at the conclusion of this study:
  - The best machine learning model is a decision tree classifier; however, the accuracy was not optimal to predict failures
  - The best parameters for a successful landing given the data analyzed appear to be:
    - Launching for the Kennedy Space Center (KSC LC-39A)
    - Using one these orbits with a payload mass lower then 8,000 kg: ES-L1, GEO, HEO and SSO
    - The VLEO orbit has a high success rate with higher payload masses (above 13,000 kg)
  - Years of experience are also required to have a successful landing, as shown when trending the success rate by year
  - A future research topic to further increase our knowledge and improve costs would be to create a model that can predict the cost of launch depending on the parameters

Thank you!