IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Zhinan Li
Jun. 20 2023

Outline

Executive Summary

- Summary of methodologies

Data Collection	Data Wrangling	Exploratory Data Analysis	Interactive visual analytics and dashboard	Predictive Analysis with Machine Learning
✓ Data collection through SpaceX API ✓ Data collection through wiki page web scraping	✓ Exploratory Data Analysis ✓ Determine Training Labels	✓ EDA with SQL ✓ EDA with Matplotlib and Seaborn	✓ Mapping with Folium lab ✓ dashboard with Plotly Dash	✓ Logistic Regression ✓ SVM ✓ Decision Tree ✓ KNN

- Summary of all results

Insights from EDA

Launch Sites Proximities Analysis

Dashboard insights

Predictive Analysis (classification)

Introduction

- **Project background and context**

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch. The goal of this project is to create a machine learning pipeline to predict if the first stage will successfully land, thus determining the launch cost.

- **Problems needed to find answers**

➢ What factors play important roles for a successful landing?
➢ how to predict if the landing will be successful or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data collection through SpaceX API
  - Data collection through wiki page web scraping
- Perform data wrangling
  - Prepare Data for Machine Learning Model Building
  - Determine Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Train Logistic Regression model, SVM, Tree Decision classifier, KNN models
  - Use GridSearchCV to conduct hyper parameter tunning

Data Collection

- Overview

SpaceX API	Web Scraping
▪ a RESTful API ▪ Make a get request ▪ Transform response content to pandas data frame ▪ Perform data cleaning	▪ Request Falcon 9 launch records from Wikipedia page ▪ Extract HTML tables with BeautifulSoup ▪ Transform data to pandas data frame ▪ Perform data cleaning

Data Collection – SpaceX API

01 Make a get request to SpaceX API url

Decode response content as JSON format and transform data into a pandas dataframe with pandas. json_normalize(). 02

03 Perform data cleaning and wrangling

Save data as .csv file for further analysis 04

GitHub URL of the completed SpaceX API calls notebook

Data Collection - Scraping

01 Scrape Falcon 9 launch records from a [Wikipedia page]

02 Parse HTML tables with BeautifulSoup and convert data to a pandas dataframe

03 Perform data cleaning and wrangling

04 Save data as .csv file for further analysis

[GitHub URL of the completed web scraping notebook]

Data Wrangling

01 Calculate the number of launches on each site

Calculate the number and occurrence of each orbit 02

03 Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column: with 1 meaning the booster successfully landed 0 means it was unsuccessful 04

GitHub URL of the completed data wrangling notebook

EDA with Data Visualization

01

Scatter plot

Visualize relationship and find correlations between different pairs of features (FlightNumber vs. PayloadMass, FlightNumber vs LaunchSite etc.)

02

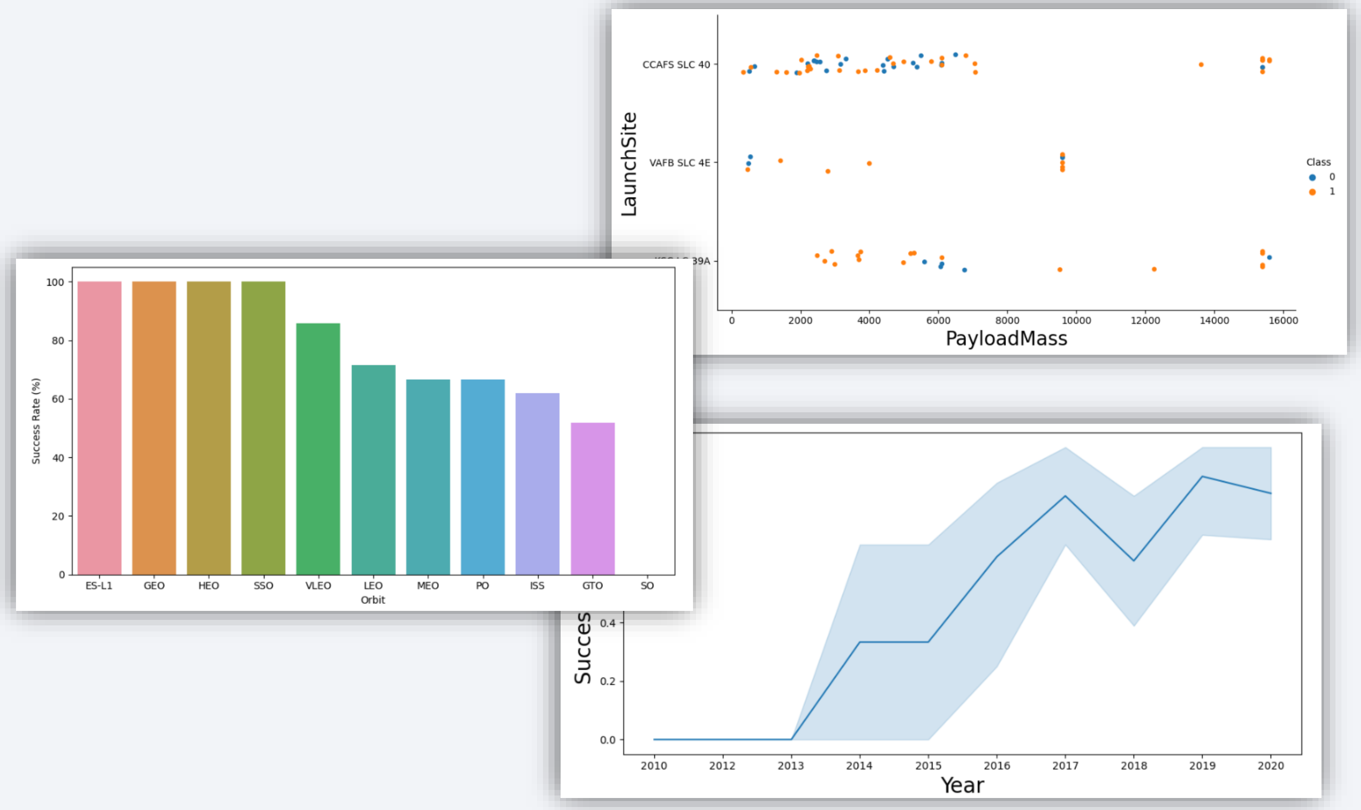
Bar chart

Compare success rate for different orbit types

03

Line chart

Visualize launch success yearly trend and check patterns



[GitHub URL of the completed EDA with data visualization notebook]

EDA with SQL

[GitHub URL of the completed EDA with SQL notebook]

1	Display the names of the unique launch sites in the space mission
2	Display 5 records where launch sites begin with the string 'CCA'
3	Display the total payload mass carried by boosters launched by NASA (CRS)
4	Display average payload mass carried by booster version F9 v1.1
5	List the date when the first successful landing outcome in ground pad was achieved.
6	List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7	List the total number of successful and failure mission outcomes
8	List the names of the booster versions which have carried the maximum payload mass. Use a subquery
9	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.
10	Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Build an Interactive Map with Folium

Step 1

Mark all launch sites as Marker objects

Step 2

Mark and visualize success/failed launches for each site using Marker Clusters

Step 3

Calculate distance between a launch site and its proximities, such as railways, highways, cities, and coastlines: find location pattern for launch site selection

GitHub URL of the completed interactive map with Folium map notebook

Build a Dashboard with Plotly Dash

01 Add a Launch Site Drop-down Input Component

02 Add a callback function to render success-pie-chart based on selected site dropdown

03 Add a Range Slider to Select Payload

04 Add a callback function to render the success-payload-scatter-chart scatter plot

[GitHub URL of the completed Plotly Dash source file]

Predictive Analysis (Classification)

Data Preparation	Data Standardization	Data Splitting	Model Training	Model Selection
Create a NumPy array from the column Class in data, by applying the method to_numpy() then assign it to the variable Y	Standardize the data in X then reassign it to the variable X	Use the function train_test_split to split the data X and Y into training and test data	✓ Logistic Regression ✓ SVM ✓ Decision Tree ✓ KNN	✓ Hyper parameter tuning for each model ✓ Select model with best test accuracy score

[GitHub URL of the completed predictive analysis notebook]

Results

Exploratory Data Analysis Results

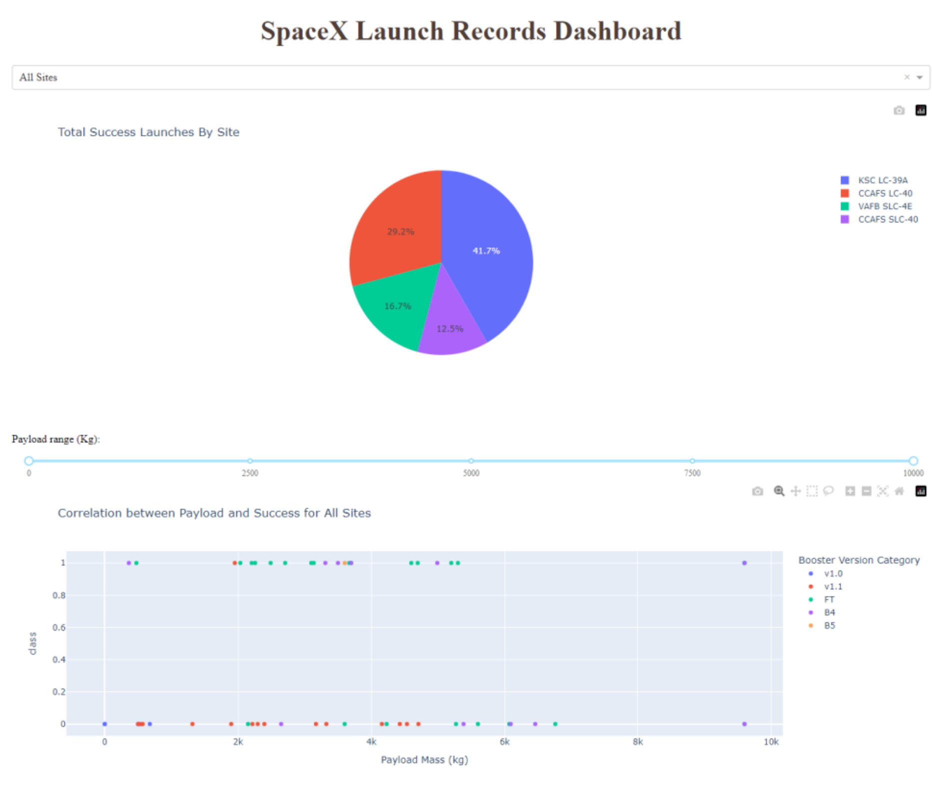
- Launch Site CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.

- success rate for VAFB SLC 4E is 100% after flight number 50. Both KSC LC 39A and CCAFS SLC 40 have a 100% success rate after flight number 80.

- orbits ES-L1, GEO, HEO, and SSO have the highest success rates at 100%

- the success rate since 2013 kept increasing till 2020

Interactive Dashboard Overview



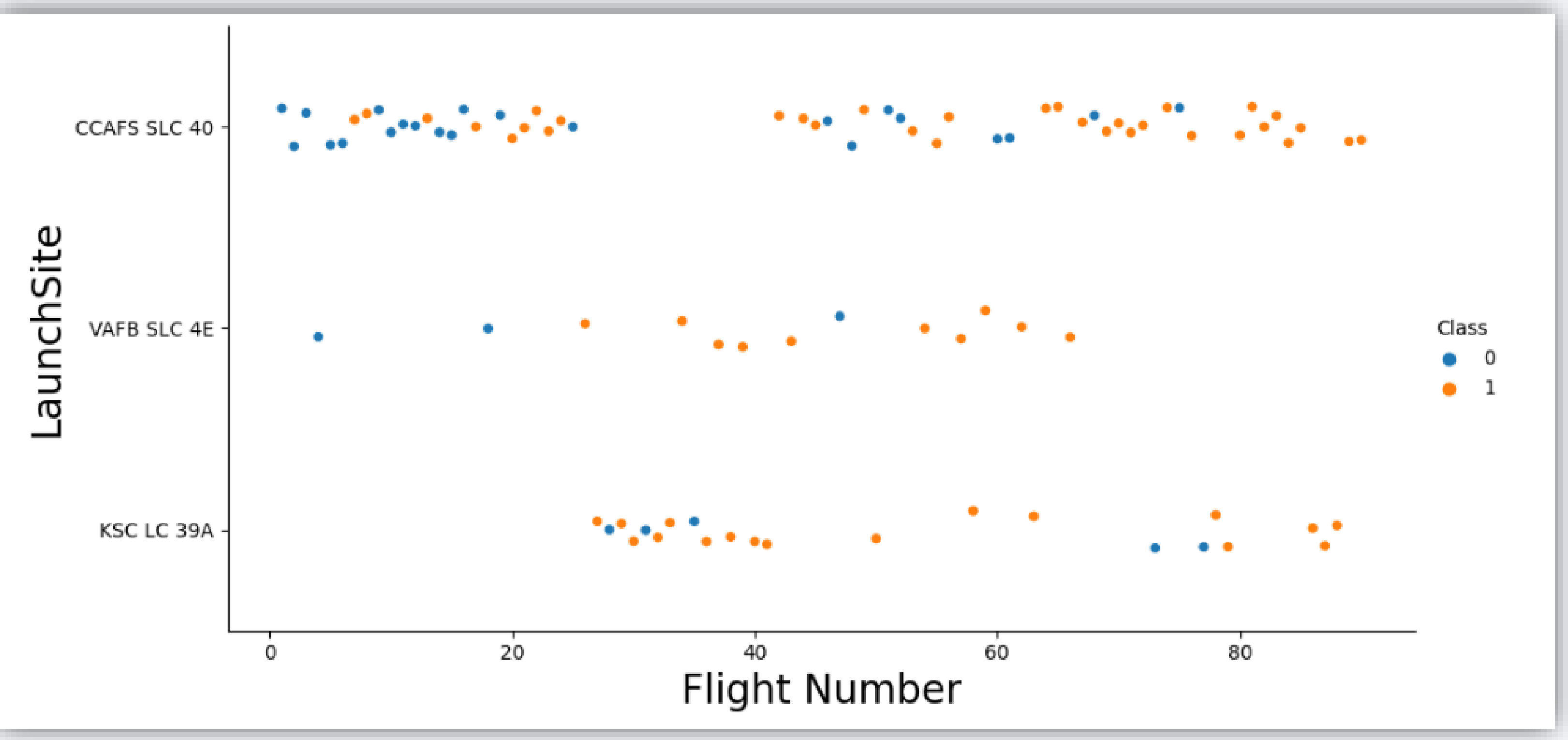
Predictive Analysis Results

all the 4 trained classifiers reached almost the same test accuracy score of around 83%

Section 2

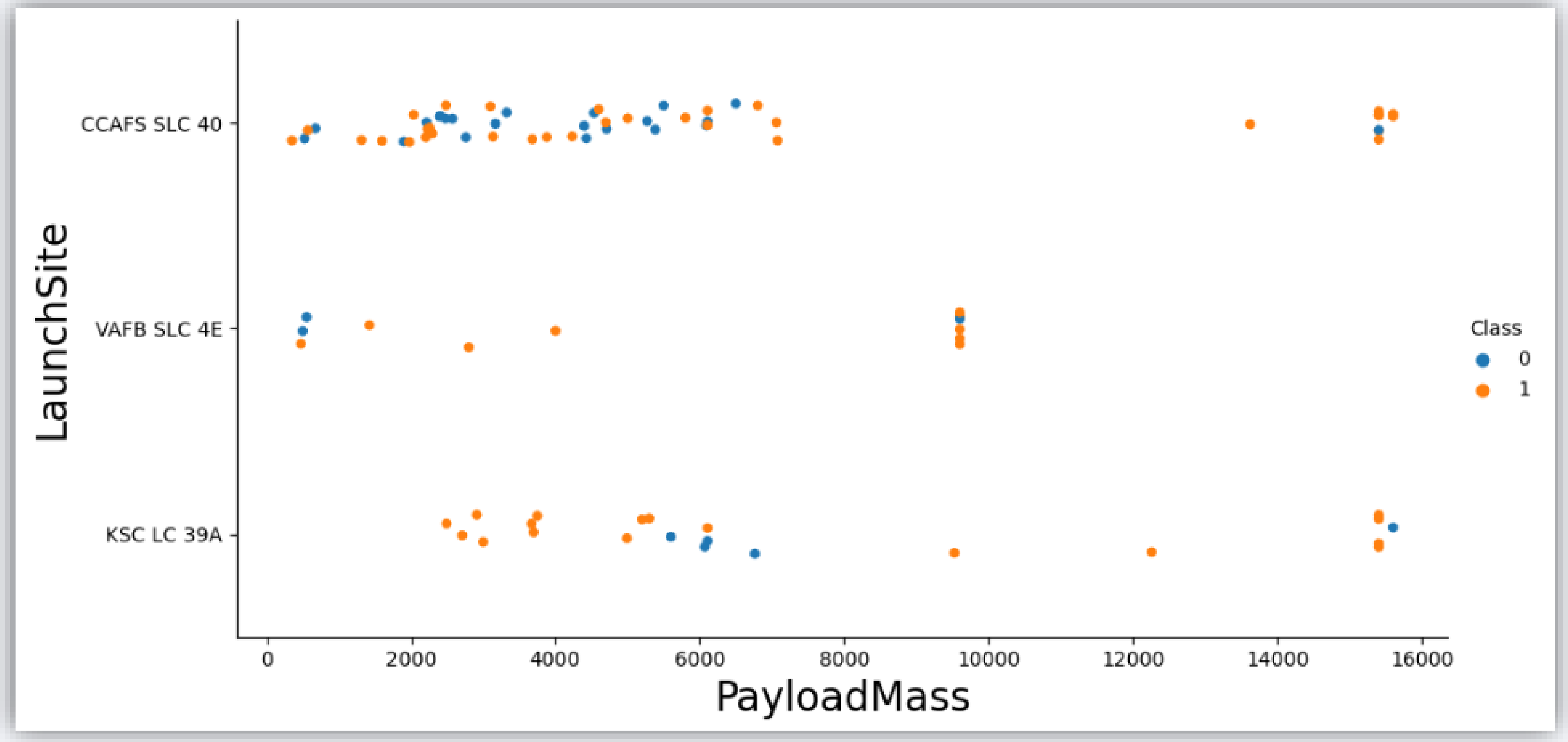
Insights drawn from EDA

Flight Number vs. Launch Site



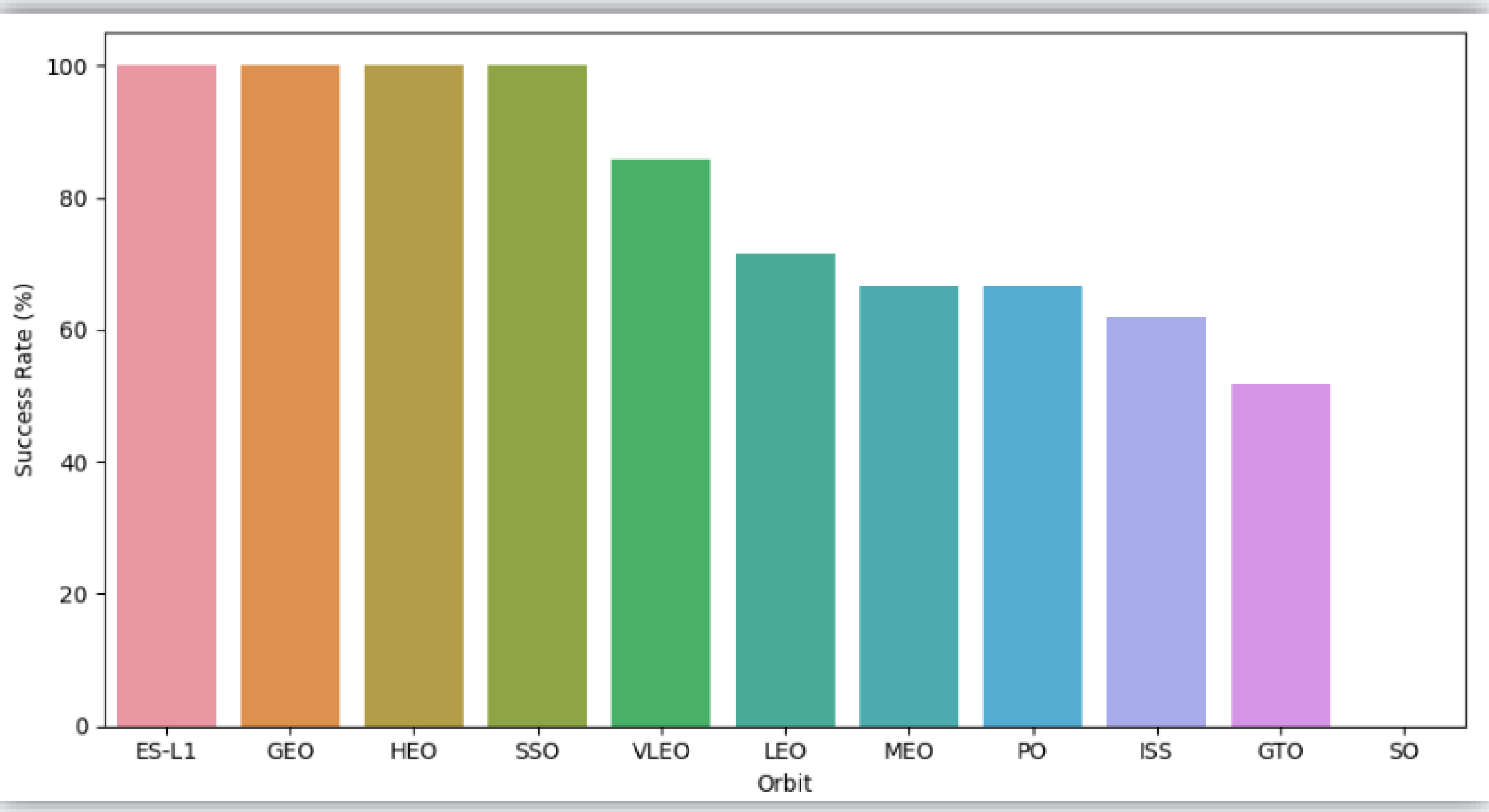
- the success rate for VAFB SLC 4E launch site is 100% after flight number 50.
- both KSC LC 39A and CCAFS SLC 40 have a 100% success rate after flight number 80.

Payload vs. Launch Site



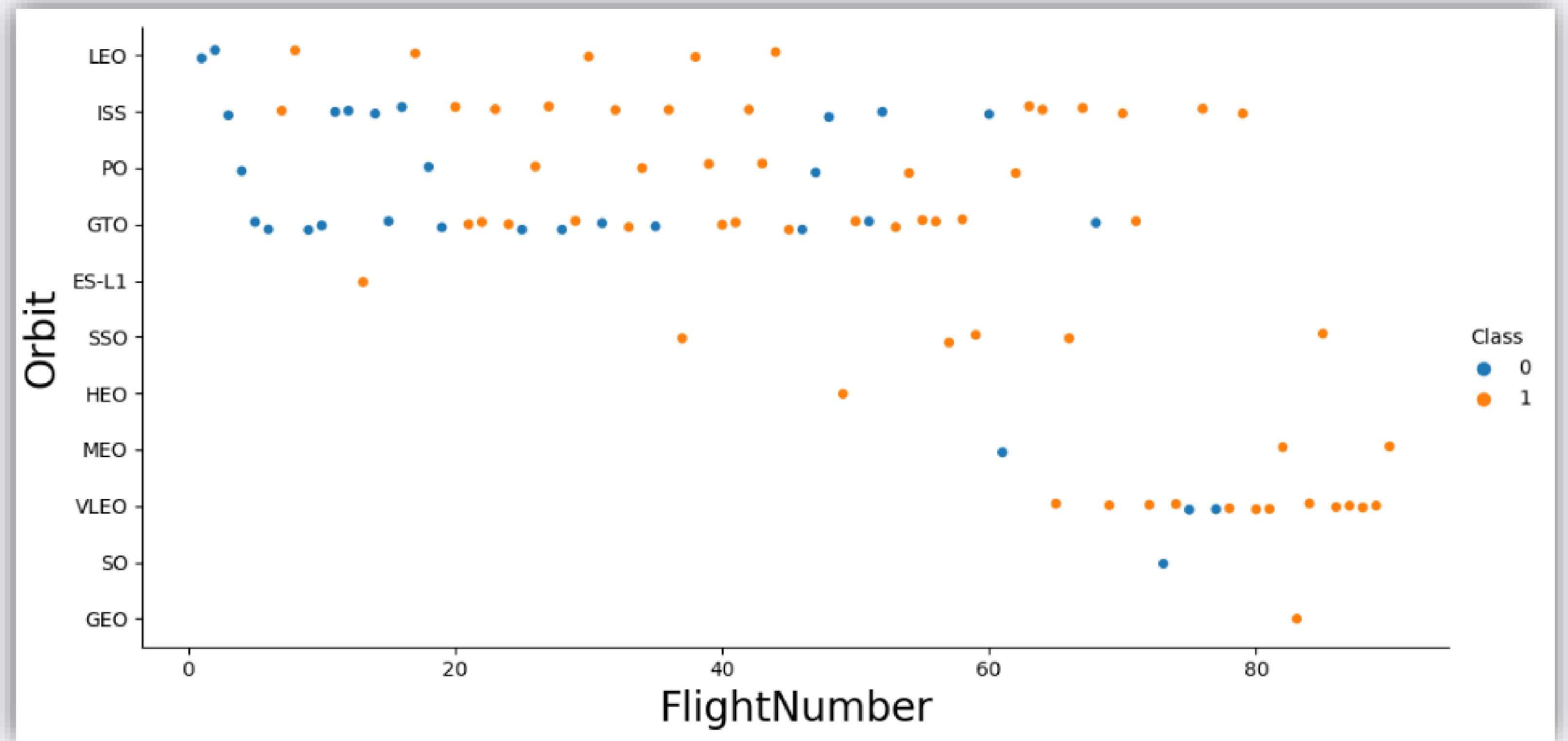
- for the VAFB-SLC launch site there are no rockets launched for heavy payload mass (greater than 10000)

Success Rate vs. Orbit Type



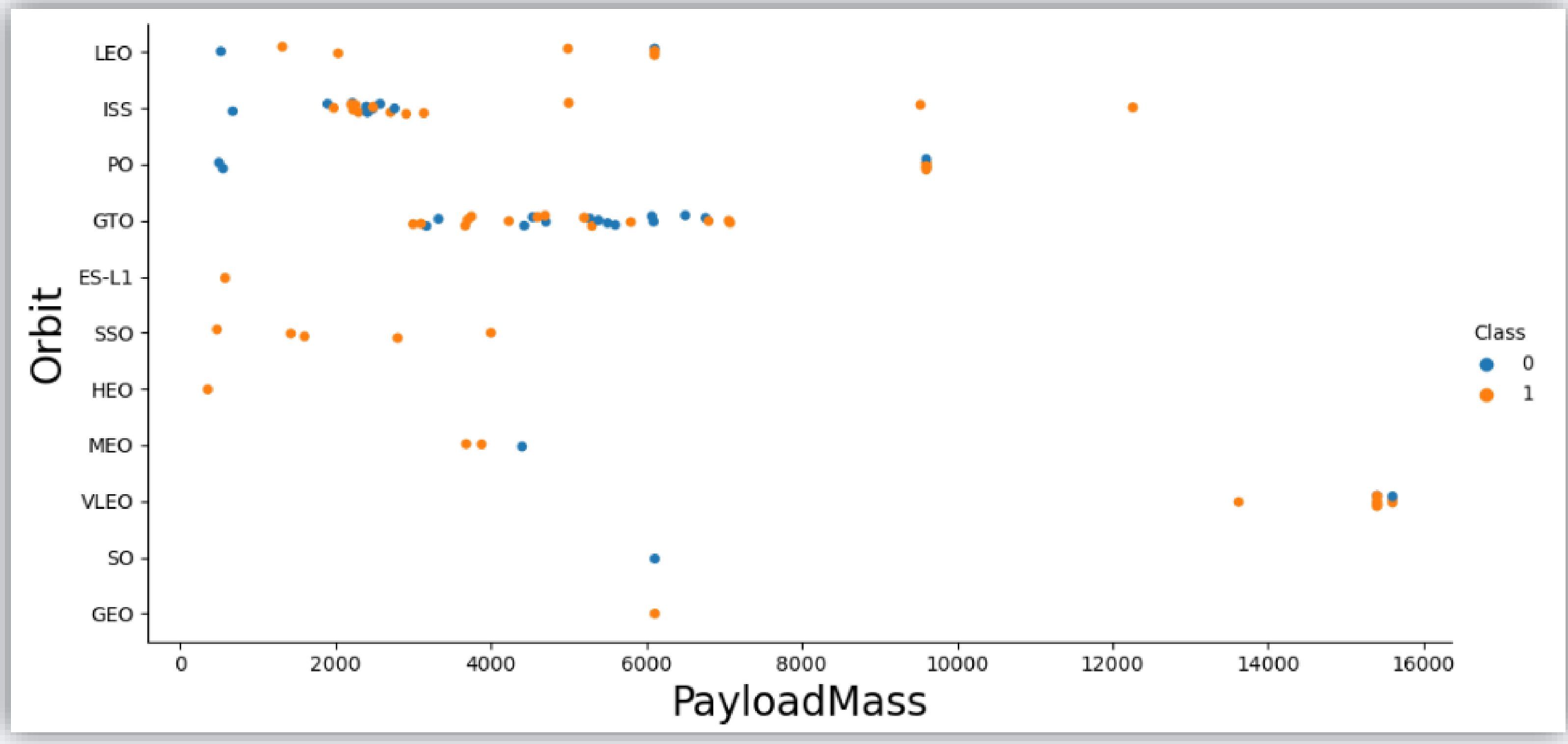
- orbits ES-L1, GEO, HEO, and SSO have the highest success rates at 100%
- SO orbit has 0% success rate
- other orbits all have around 60% success rate

Flight Number vs. Orbit Type



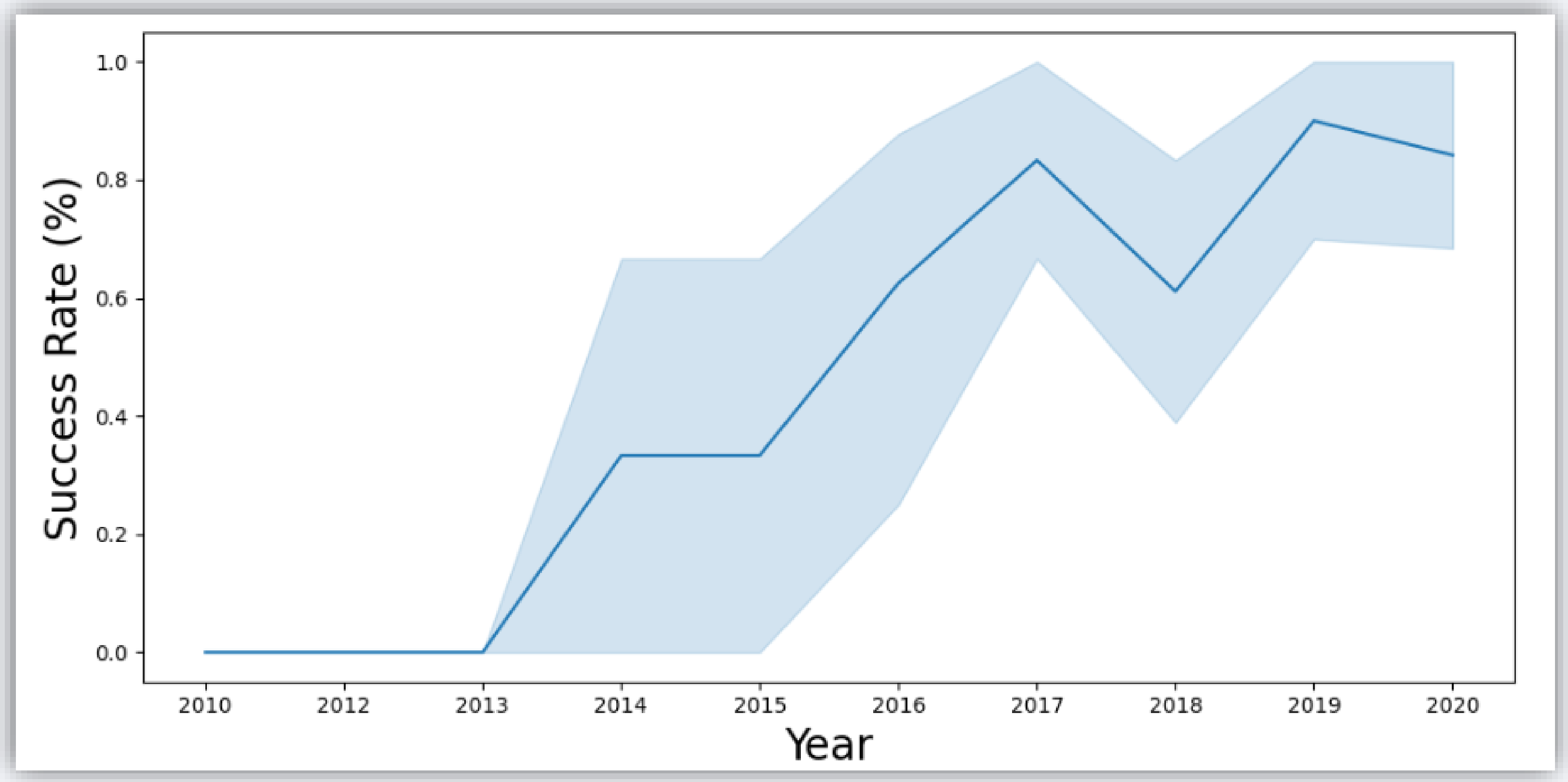
- in the LEO orbit the Success appears related to the number of flights
- there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- ➢ With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
- ➢ However for GTO we cannot distinguish this well as both positive and negative landing are there.

Launch Success Yearly Trend



- the success rate kept increasing since 2013 till 2020

All Launch Site Names

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

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

- Use SELECT DISTINCT statement to return only the unique launch sites
- There are 4 different launch sites

Launch Site Names Begin with 'CCA'

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

- **Used LIKE key word with % wildcard in WHERE clause to filter the records**
- **First 5 records all have failed landing outcomes**

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) TOTAL_PAYLOAD_MASS_KG FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

TOTAL_PAYLOAD_MASS_KG
45596.0

- Use SUM() aggregate function to return and display the total sum of payload
- Total payload mass carried by boosters launched by NASA (CRS) is 45,596 kg

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AVG_PAYLOAD_MASS_KG FROM SPACEXTBL WHERE BOOSTER_VERSION LIKE 'F9 v1.1%';
```

AVG_PAYLOAD_MASS_KG
2534.666666666665

- **Use AVG() aggregate function to return and display the average payload mass**
- **The average payload mass carried by booster version F9 v1.1 is around 2534.67 kg**

First Successful Ground Landing Date

```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME = 'Success (ground pad)';
```

MIN(DATE)
01/08/2018

- Use MIN() aggregate function to return and display the date when the first successful landing outcome in ground pad was achieved.
- The first successful ground pad landing was on 01.08.2018

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD, CUSTOMER, PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE LANDING_OUTCOME = 'Success (drone ship)'
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000
```

Booster_Version	Payload	Customer	PAYLOAD_MASS__KG_
F9 FT B1022	JCSAT-14	SKY Perfect JSAT Group	4696.0
F9 FT B1026	JCSAT-16	SKY Perfect JSAT Group	4600.0
F9 FT B1021.2	SES-10	SES	5300.0
F9 FT B1031.2	SES-11 / EchoStar 105	SES EchoStar	5200.0

- **Use multiple conditions in WHERE clause to filter data as required**
- **There are 4 records that meet the conditions**

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(MISSION_OUTCOME) TOTAL_NUMBER FROM SPACEXTBL GROUP BY MISSION_OUTCOME;
```

Mission_Outcome	TOTAL_NUMBER
None	0
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

- **Use COUNT() aggregate function together with GROUP BY statement to return total number of mission outcomes for each group**
- **There are 1 failure in flight, 99 successes and 1 success with unclear payload status.**

Boosters Carried Maximum Payload

```
%%sql
SELECT DISTINCT BOOSTER_VERSION, PAYLOAD, CUSTOMER, PAYLOAD_MASS__KG_ FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

Booster_Version	Payload	Customer	PAYLOAD_MASS__KG_
F9 B5 B1048.4	Starlink 1 v1.0, SpaceX CRS-19	SpaceX	15600.0
F9 B5 B1049.4	Starlink 2 v1.0, Crew Dragon in-flight abort test	SpaceX	15600.0
F9 B5 B1051.3	Starlink 3 v1.0, Starlink 4 v1.0	SpaceX	15600.0
F9 B5 B1056.4	Starlink 4 v1.0, SpaceX CRS-20	SpaceX	15600.0
F9 B5 B1048.5	Starlink 5 v1.0, Starlink 6 v1.0	SpaceX	15600.0
F9 B5 B1051.4	Starlink 6 v1.0, Crew Dragon Demo-2	SpaceX	15600.0
F9 B5 B1049.5	Starlink 7 v1.0, Starlink 8 v1.0	SpaceX, Planet Labs	15600.0
F9 B5 B1060.2	Starlink 11 v1.0, Starlink 12 v1.0	SpaceX	15600.0
F9 B5 B1058.3	Starlink 12 v1.0, Starlink 13 v1.0	SpaceX	15600.0
F9 B5 B1051.6	Starlink 13 v1.0, Starlink 14 v1.0	SpaceX	15600.0
F9 B5 B1060.3	Starlink 14 v1.0, GPS III-04	SpaceX	15600.0
F9 B5 B1049.7	Starlink 15 v1.0, SpaceX CRS-21	SpaceX	15600.0

- **Use a subquery to get the max payload and used it in outer query to list all the boosters that have carried the max payload**
- **The max payload mass is 15600 kg**

2015 Launch Records

```
%%sql
SELECT SUBSTR(DATE, 4, 2) MONTH, SUBSTR(DATE, 7, 4) YEAR, LANDING_OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE SUBSTR(DATE, 7, 4) = '2015' AND LANDING_OUTCOME = 'Failure (drone ship)';
```

MONTH	YEAR	Landing_Outcome	Booster_Version	Launch_Site
10	2015	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- **Use SUBSTR() scalar function to get the month and year from the DATE column, and use multiple conditions in WHERE clause to filter data**
- **In 2015 there are 2 launch failures, in Apr. and Oct. respectively, both are in CCAFS LC-40 launch site**

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

```
%%sql
SELECT LANDING_OUTCOME, COUNT(*) TOTAL FROM SPACEXTBL
WHERE LANDING_OUTCOME LIKE 'Success%'
AND DATE BETWEEN '04-06-2010' AND '20-03-2017'
GROUP BY LANDING_OUTCOME
ORDER BY 2 DESC;
```

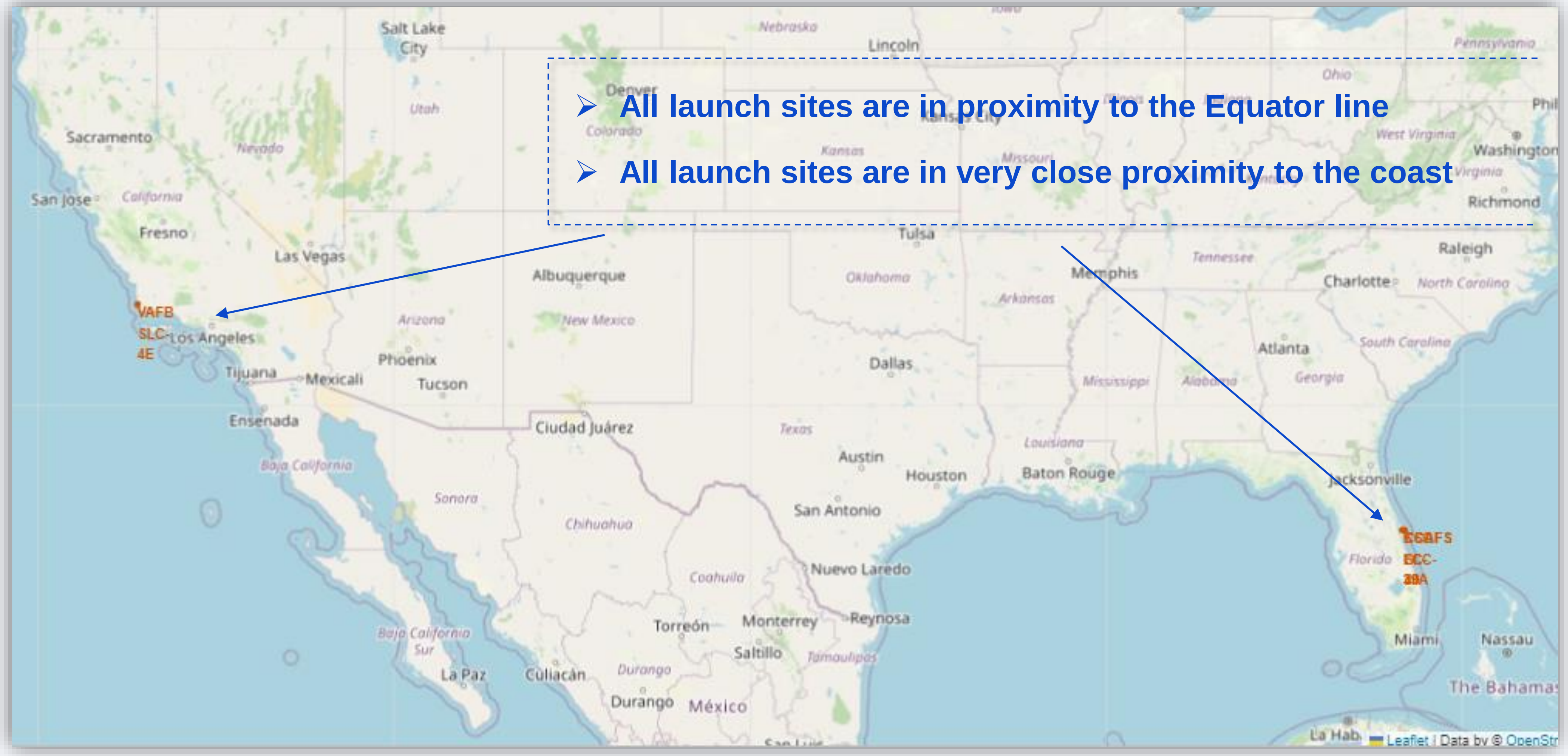
Landing_Outcome	TOTAL
Success	20
Success (drone ship)	8
Success (ground pad)	7

- **Use GROUP BY and ORDER BY to rank the count of landing outcomes between the date 2017-03-20 and 2010-06-04 in descending order**
- **During this period, there are 20 successful landing, 8 successful drone ship landing and 7 successful ground pad landing**

Section 3

Launch Sites Proximities Analysis

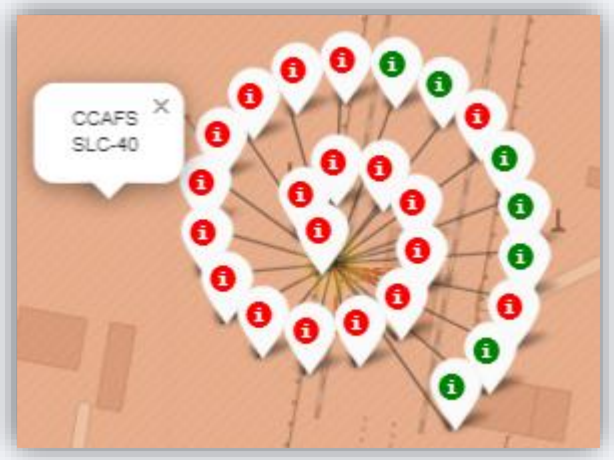
All launch sites on a map



➢ **All launch sites are in proximity to the Equator line**

➢ **All launch sites are in very close proximity to the coast**

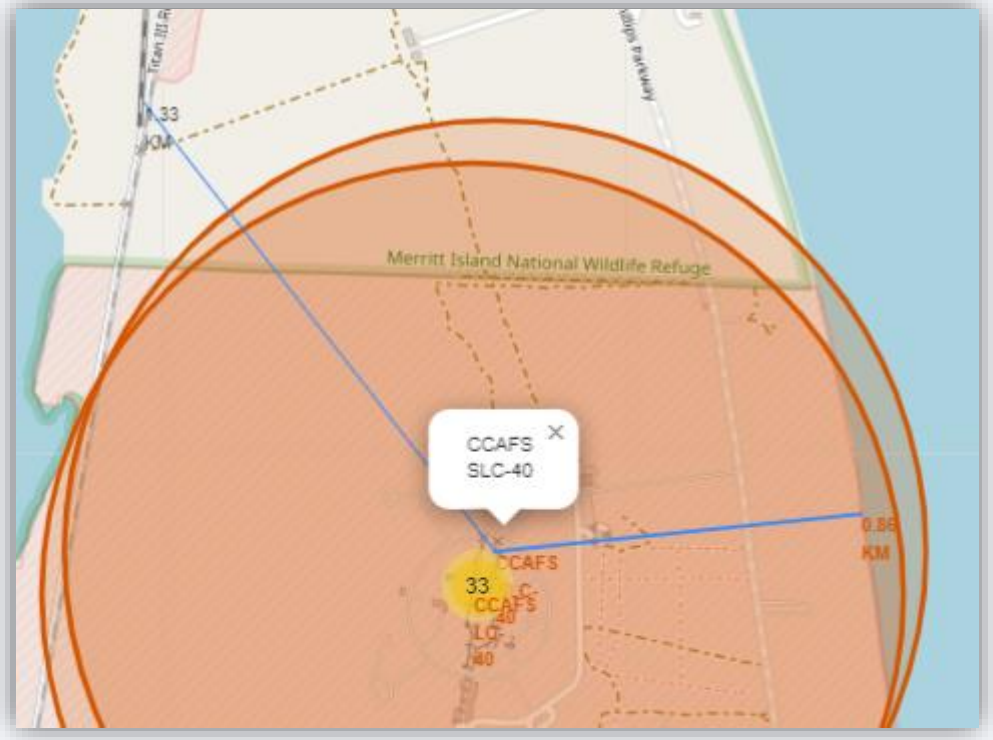
Success/failed launches for each site on the map



- ➢ In the Eastern coast, Launch site KSC LC-39A has relatively higher success rates than CCAFS SLC-40 and CCAFS LC-40.

- ➢ In the West Coast, Launch site VAFB SLC-4E has a success rate of 4/10

distances between a launch site to its proximities

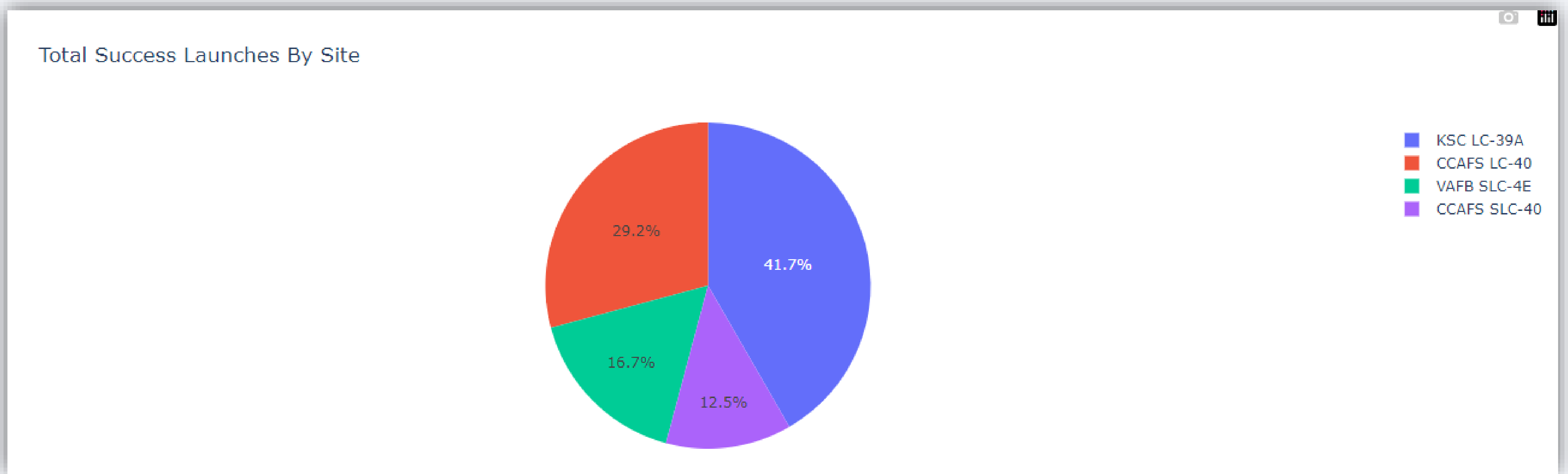


- Launch site CCAFS SLC-40 is about 0.86km away from its proximate coastline
- Launch site CCAFS SLC-40 is about 1.33km away from its proximate railway

Section 4

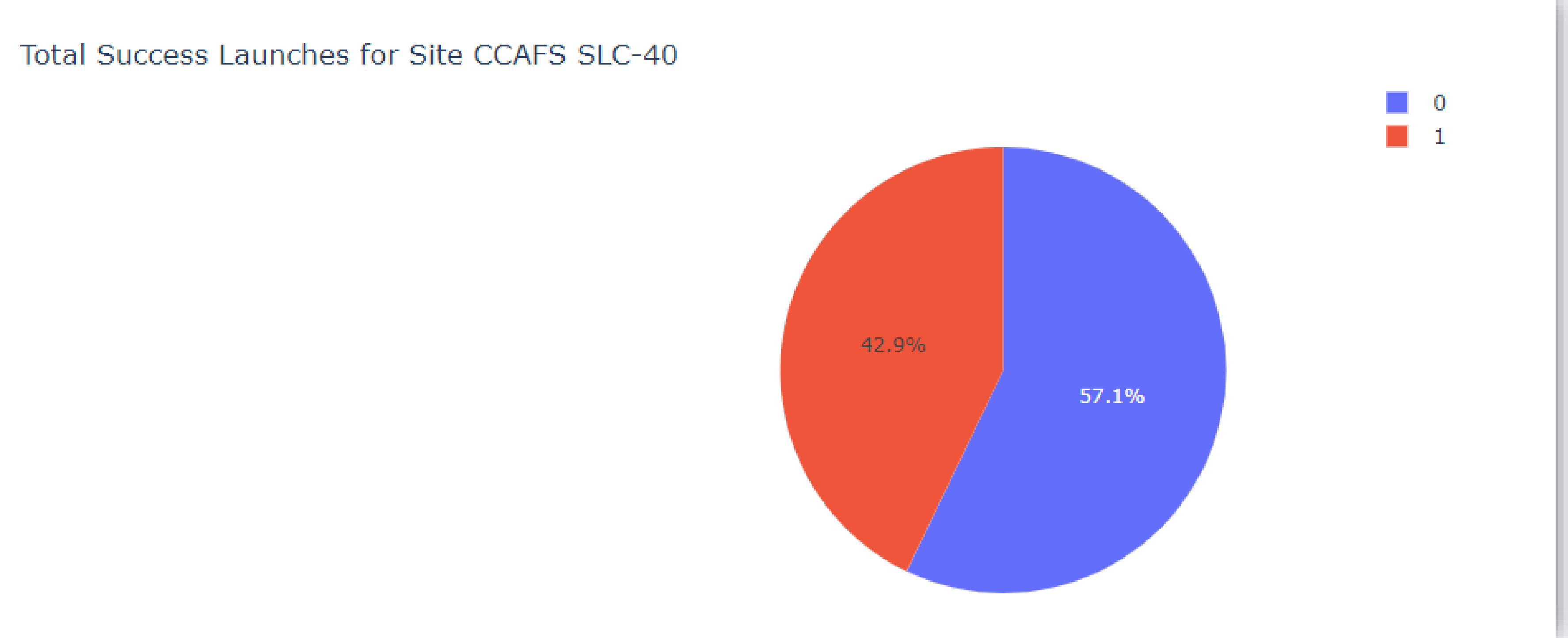
Build a Dashboard with Plotly Dash

Launch success count for all sites



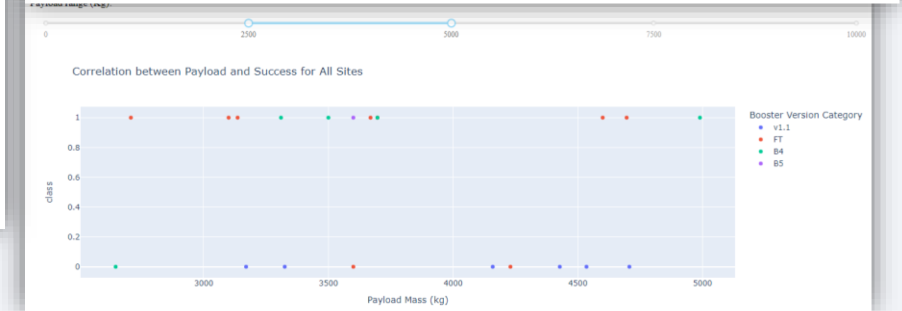
- Launch site KSC LC-39A accounts for highest launch success counts at 41.7% out of total counts
- Success counts for CCAFS LC-40 and VAFB SLC-4E take 29.2% and 16.7% respectively out of total
- launch site CCAFS SLC-40 has the lowest success counts of 12.5% out of total

launch site with highest launch success ratio



➢ **Launch site KSC LC-39A has the highest launch success rate at 42.9**

Payload vs. Launch Outcome scatter plot for all sites

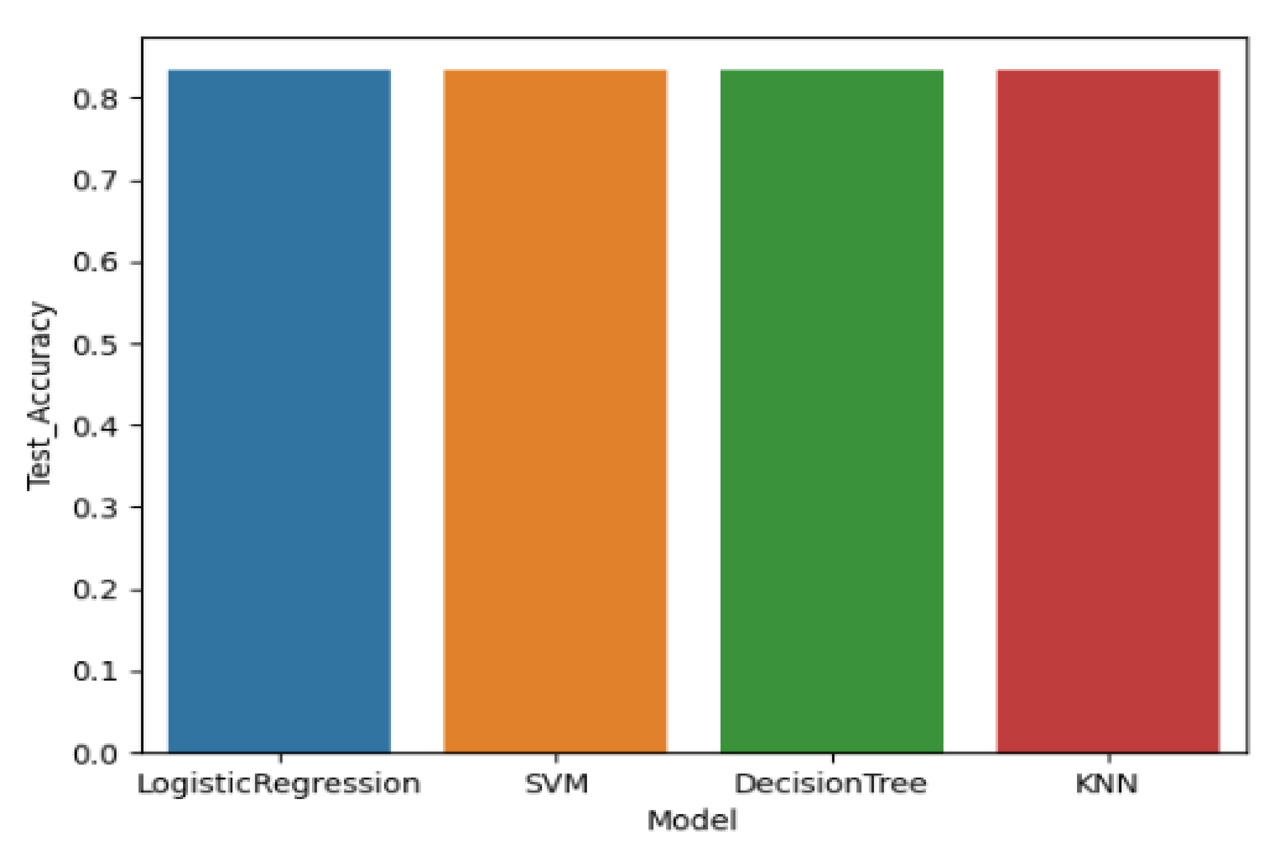


- ➢ **Booster version FT has the largest success rate from a payload mass of > 2000 kg**
- ➢ **V1.0 can take heaviest payload mass**
- ➢ **The success landing happens mostly at payload mass ranging from 2000 kg to 6000 kg**

Section 5

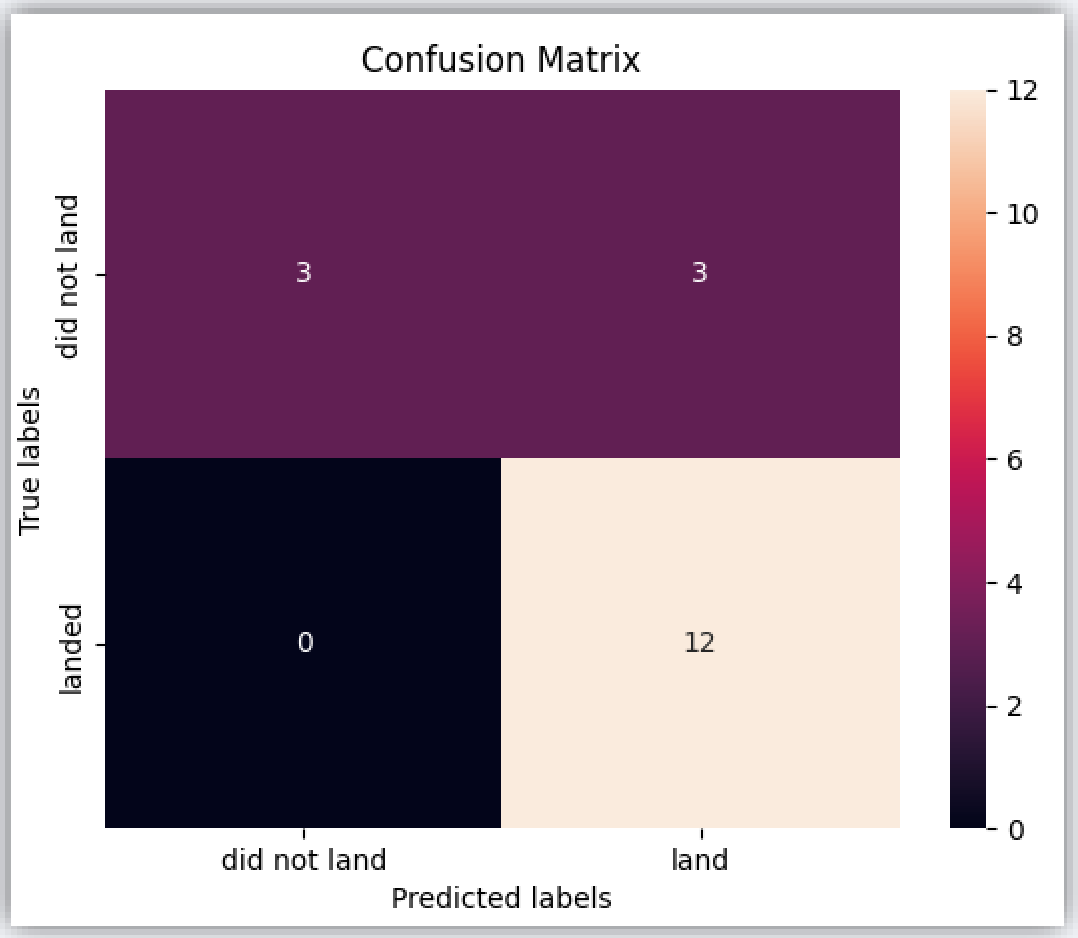
Predictive Analysis (Classification)

Classification Accuracy



- ALL models perform equally on test data with a test accuracy score at 83.33%

Confusion Matrix



- All of 4 models have the same confusion matrix
- They can distinguish between the different classes. The major problem is false positives.

Conclusions

- Launch Site CCAFS LC-40 has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%.
- As flight number increases, the success rate also increases. The success rate kept increasing since 2013 till 2020.
- orbits ES-L1, GEO, HEO, and SSO have the highest success rates at 100%.
- With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
- With hyper parameter tunning, ALL 4 trained classifiers perform equally on test data with a test accuracy score at 83.33%.

Appendix

GitHub Repository URL for this project

Thank you!