IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joel O'Brien
6 March 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Insights and outlook

Executive Summary

In order to provide an emerging company SpaceY with a machine learning model to predict launch success, data from SpaceX was collected from the SpaceX REST API server and the Wikipedia page. Get requests and webscraping was employed to extract the data, and Pandas was used to load the data into dataframes. Exploratory data analysis was performed by creating plots to identify major relationships, and SQL queries were employed to supplement this. An interactive Folium map was created to investigate the impact of geospatial factors on launch success, and a Plotly Dash was used to gain an overview of success rate per site and payload mass range. The site KSC LC-39A was the most successful, and the most promising orbits were ES-L1, GEO, HEO and SSO. Four predictive models were trained and evaluated, with a decision tree model yielding superior predictions with an accuracy of 94.4 %.

Introduction

- A competitor to SpaceX featuring reusable rocket engines requires a model to predict launch successes. There is plentiful data available on the SpaceX REST API and the Wikipedia page to provide a training data set for such a model.
- In this project, the data is collected and wrangled to assemble a workable dataset where exploratory data analysis can be performed, an interactive map and a dashboard created and eventually a trained model.
- The problems which require answering involve factors behind the success rate of rocket launches such:
  - Which payload mass range is most likely to succeed?
  - Which orbit types have the highest success rate?
  - Is there any relationship between launch site location and success rate?
  - What is the time frame for a company to progress from total launch failures to a high level of success rate?

Section 1

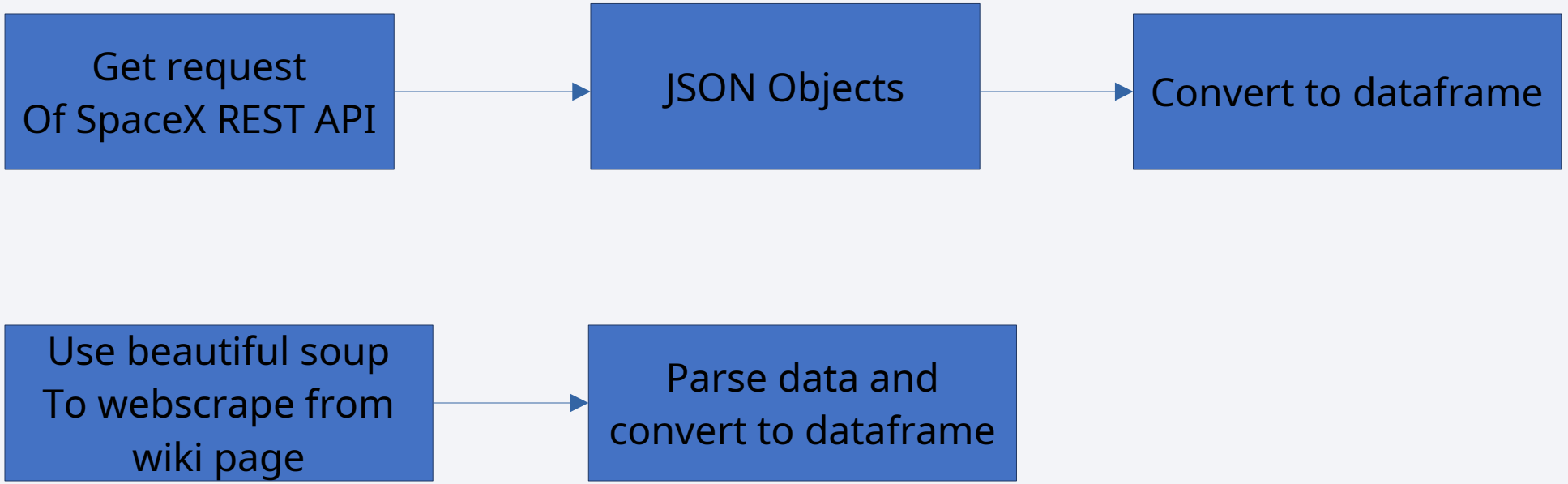
Methodology

Methodology

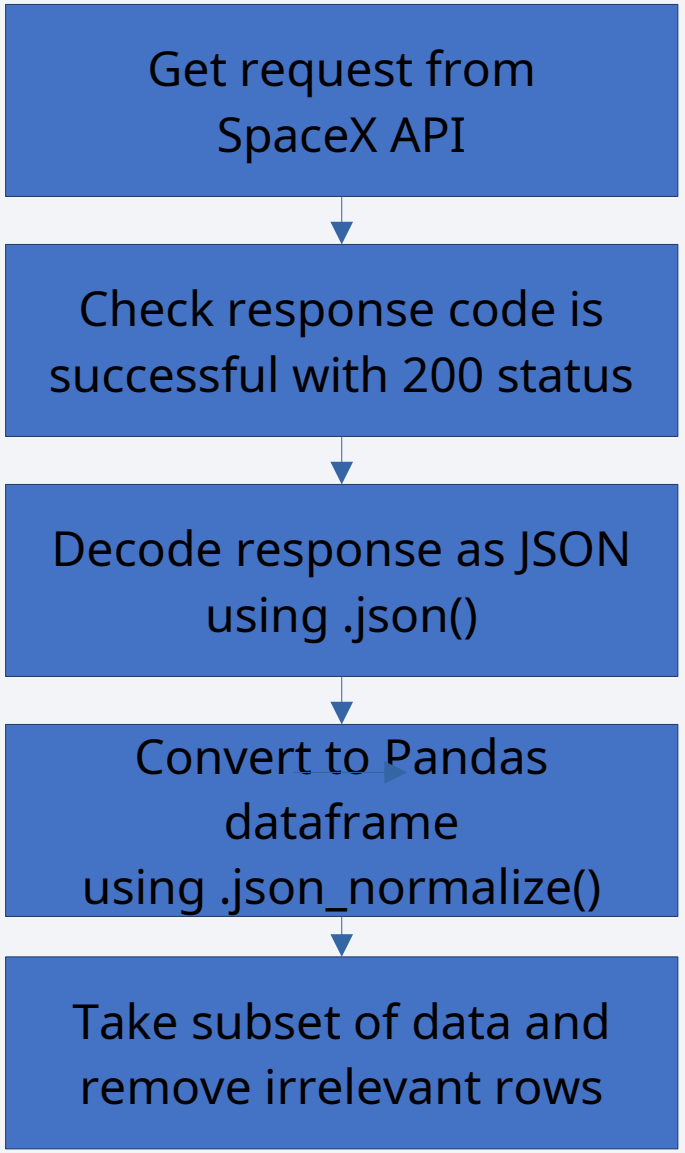
Executive Summary

- Data collection methodology:
  - Data was collected using get request via the SpaceX REST API and using BeautifulSoup to webscrape the SpaceX wikipedia page.
- Perform data wrangling
  - Data wrangling was performed using Pandas dataframes.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

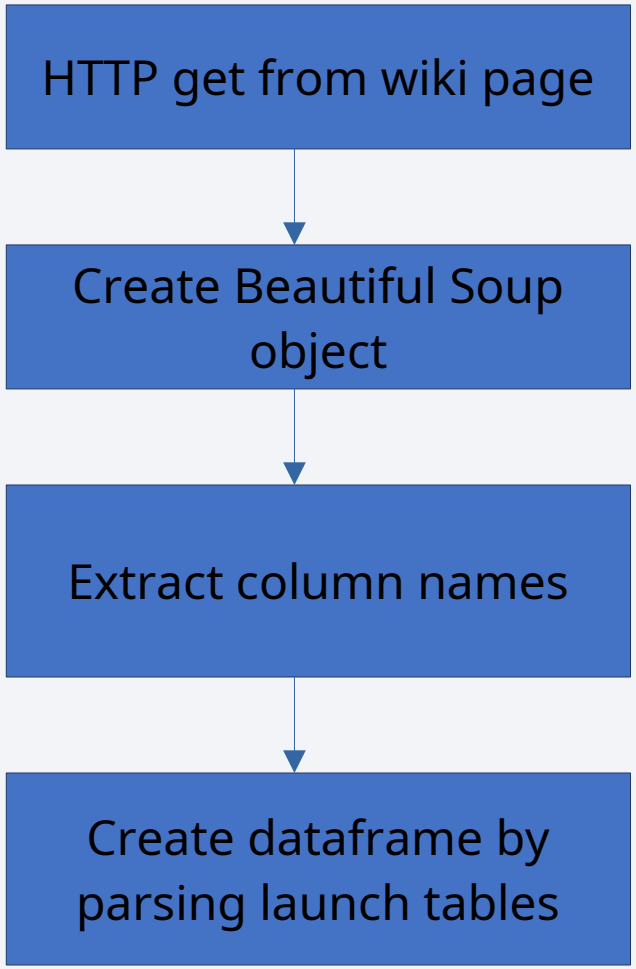


Data Collection – SpaceX API



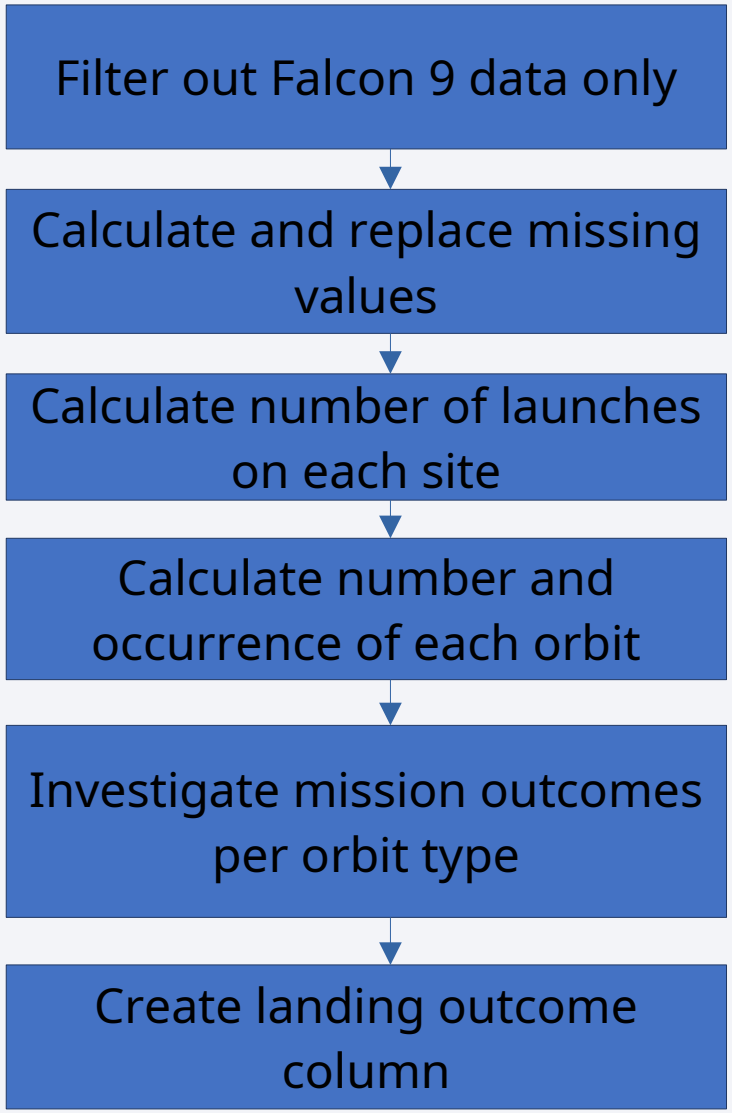
Data Collection - GitHub

Data Collection - Scraping



Webscraping - GitHub

Data Wrangling



Data Wrangling - GitHub

EDA with Data Visualization

The following charts were produced to gain an overview of the data and identify relationships between different variables:

- Flight number vs. payload mass.
- Launch site vs. flight number.
- Payload mass vs. launch site.
- Orbit type vs. success rate.
- Flight number vs. orbit type.
- Payload mass vs. orbit type.
- Success rate vs. year.

EDA Viz - GitHub

EDA with SQL

The SQL queries which were performed can be summarised as:

- Select all unique launch sites.
- Display 5 records of launch sites beginning with ‘CCA’.
- Display total payload mass of NASA flights.
- Display average payload mass carried by F9 v.1.1 boosters.
- List date of first successful ground pad landing.
- List names of boosters involved in successful drone ship landings with payload mass between 4000-6000 kg.
- List total number of successful and unsuccessful mission outcomes.
- List booster versions which carried maximum payload mass.
- Display records for failed launches by months in 2015.
- Rank count of successful landings between 04/06/2010 – 20/03/2017 in descending order.

EDA SQL - GitHub

Build an Interactive Map with Folium

The Folium interactive maps contained the following elements:

- Launch site location marker, to identify where launches took place.
- Number of launches label over location marker, to show which sites were used the most.
- Colour-coded landing outcome marker, for easy identification of which sites had the greatest number of successful landings.
- Line showing proximity to nearest coastline, to highlight safety profile of launch site in case of launch anomaly.
- Line showing proximity to nearest railway/highway, to show ease of logistics of rocket delivery.
- Line showing proximity to Orlando, to highlight safety distance from major cities.

Folium - GitHub

Build a Dashboard with Plotly Dash

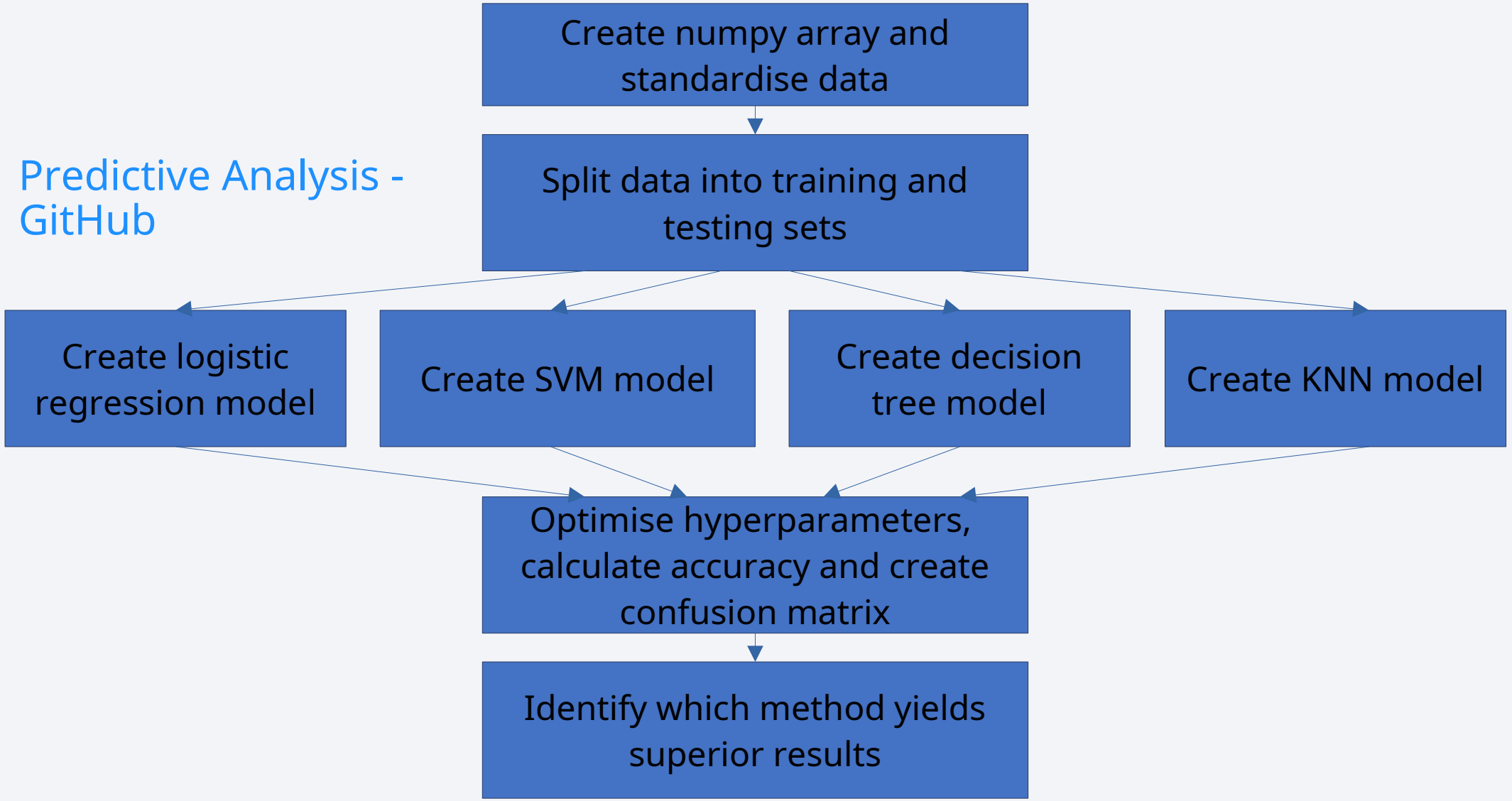
The following elements were added to the Plotly Dashboard:

- A drop-down list to select ‘all sites’ or a particular site.
- A pie chart showing the proportion of successful and failed launches.
- A slider to select payload mass range.
- A scatter plot of launch outcome vs payload mass.

Plotly Dash - GitHub

Predictive Analysis (Classification)

Predictive Analysis - GitHub



Results

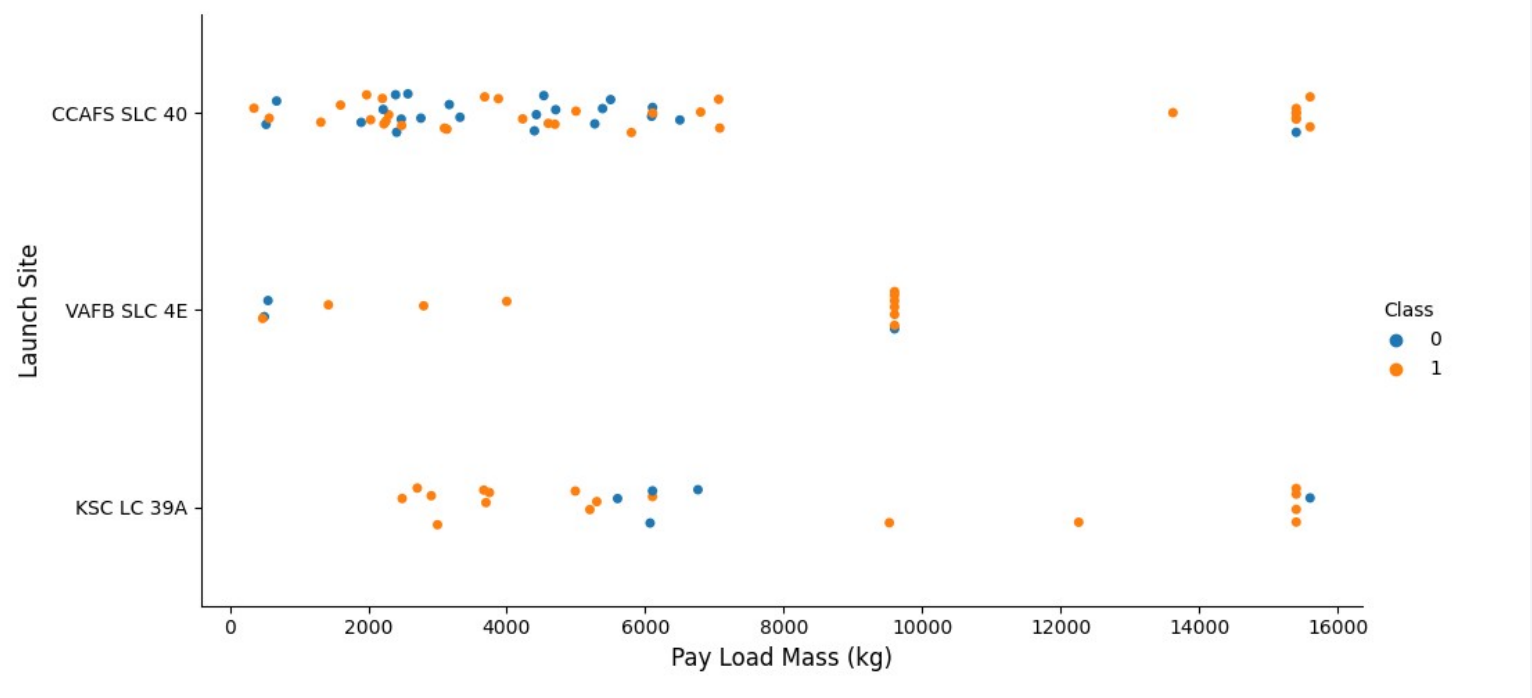
- Exploratory data analysis results: there is a strong relationship between launch success outcome and payload mass and flight number.
- Interactive analytics demo in screenshots: Folium maps indicate which launch sites are most popular and the success rate of each. The proximity to major cities and infrastructure is also shown.
- Predictive analysis results: four machine learning models were employed and the superior model was the decision tree model.

Section 2

Insights drawn from EDA

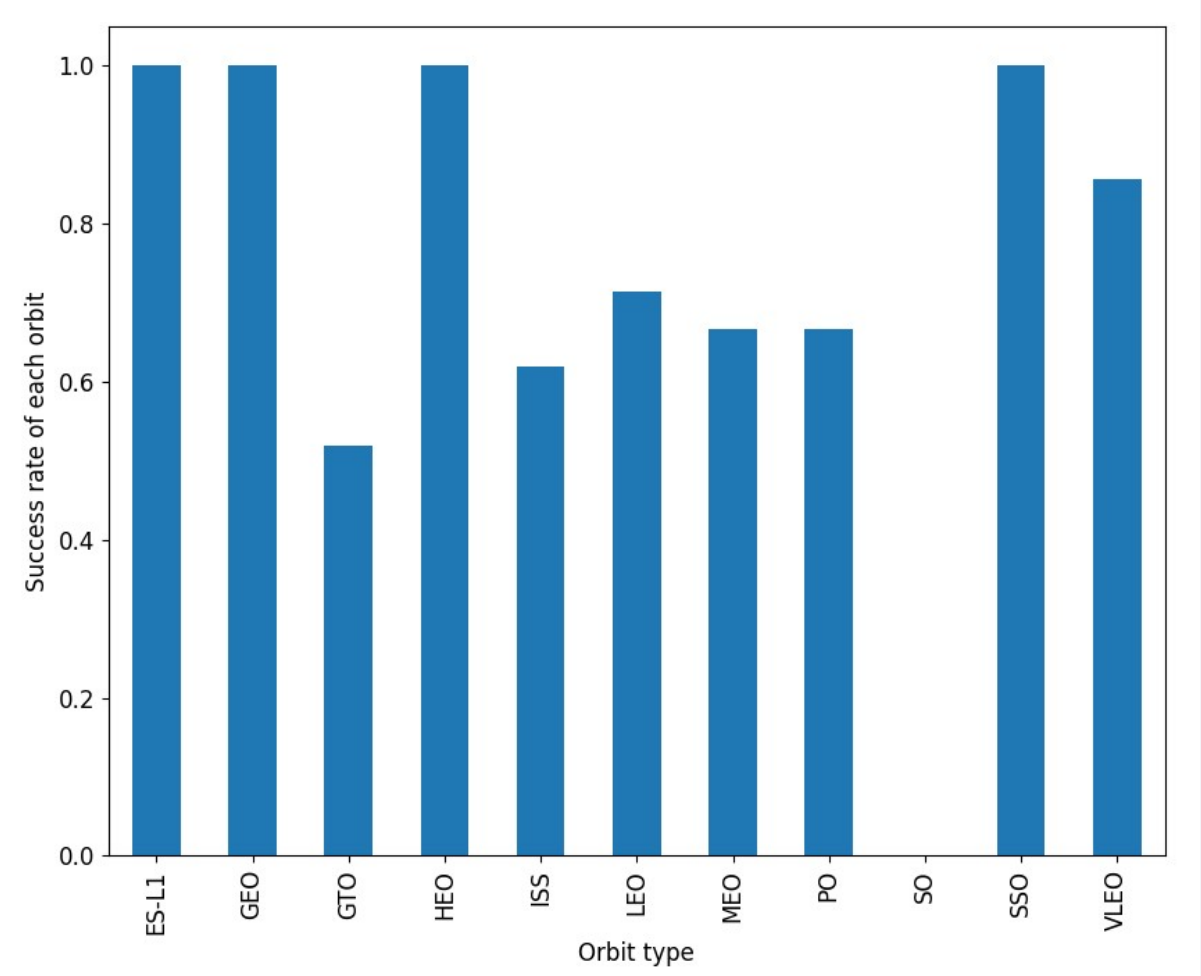
Payload vs. Launch Site

- Launch site CCAFS SLC 40 hosted the greatest number of launches with payloads from 0 kg to 7000 kg. A small number of mostly successful launches occured for higher payloads of approx. 15000kg.
- Other launch sites had mostly successful launches where VAFB SLC 4E for payloads between 0 kg and 10000kg and KSC LC 39A for payloads between 2000 kg and 16000 kg.



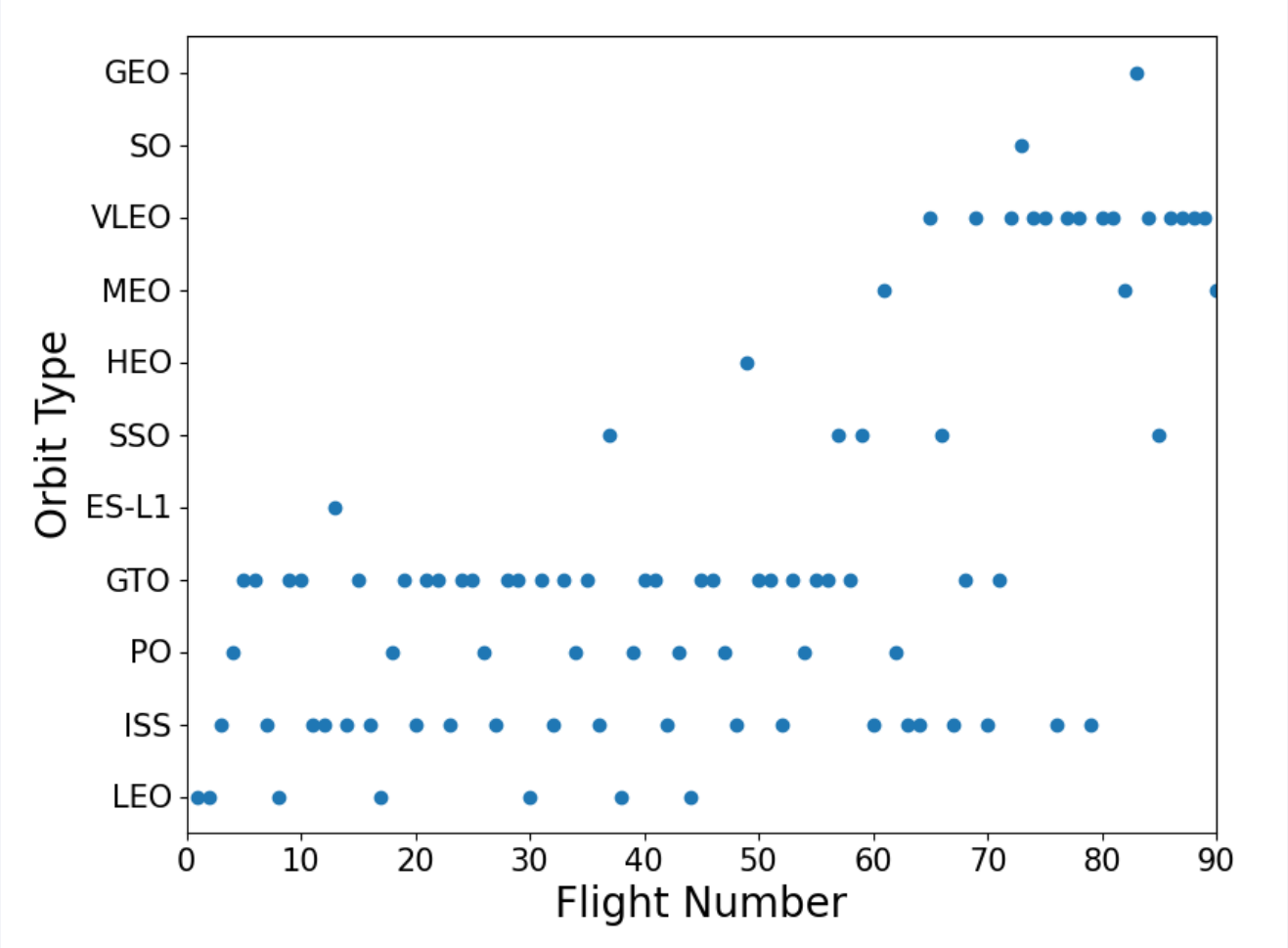
Success Rate vs. Orbit Type

- Orbit tyoes ES-L1, GEO, HEO and SSO had a 100% launch success rate.
- The SO orbit had the lowest success rate of 0%.
- Other orbit types had a success rate of approximately 60 %.



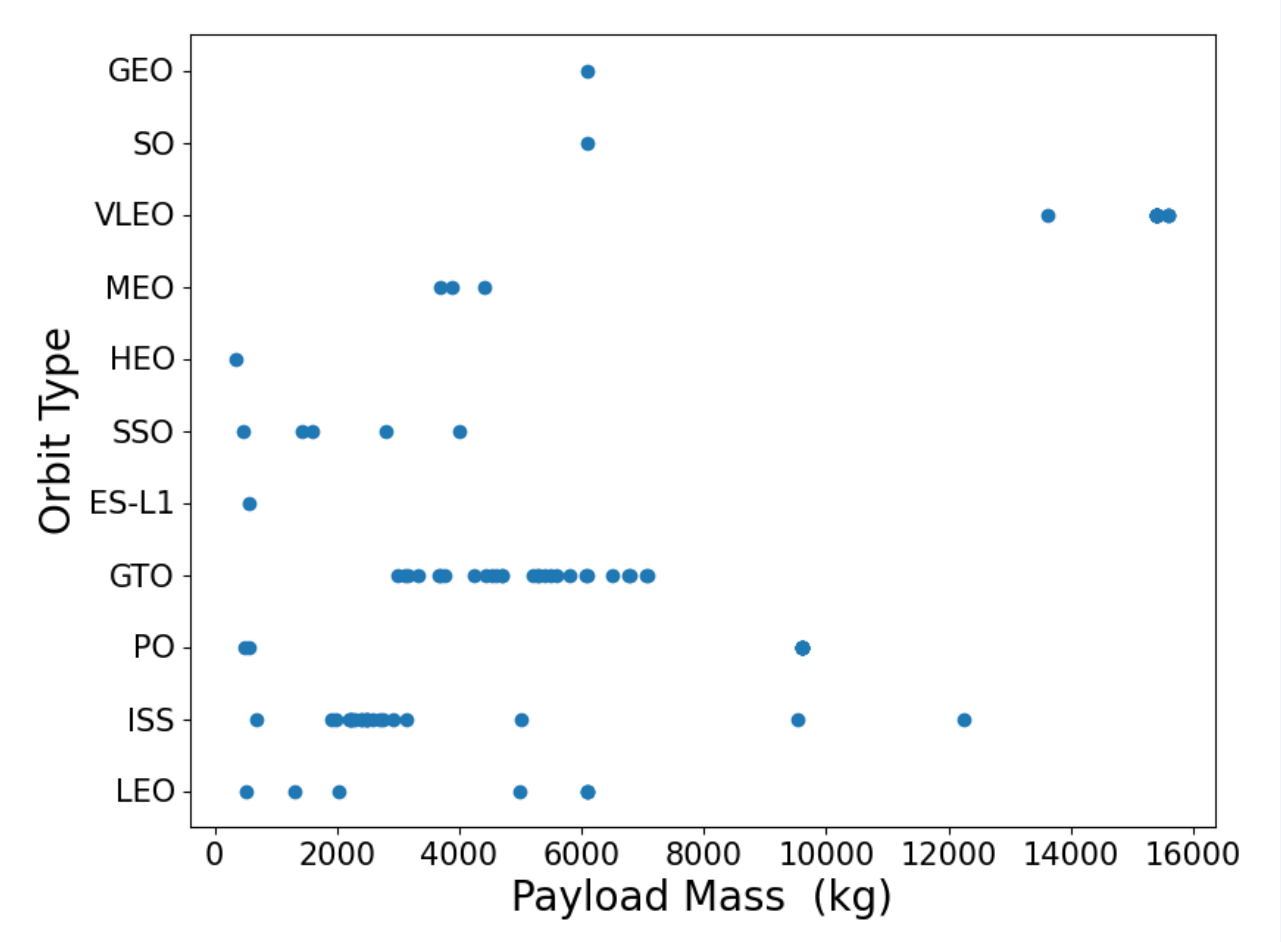
Flight Number vs. Orbit Type

- Orbit types LEO, ISS, PO and GTO were most common for early flight numbers.
- MEO, VLEO, SO and GEO were the most common orbit types for later flight numbers.
- ISS had the greatest range of launches throughout all flight numbers.



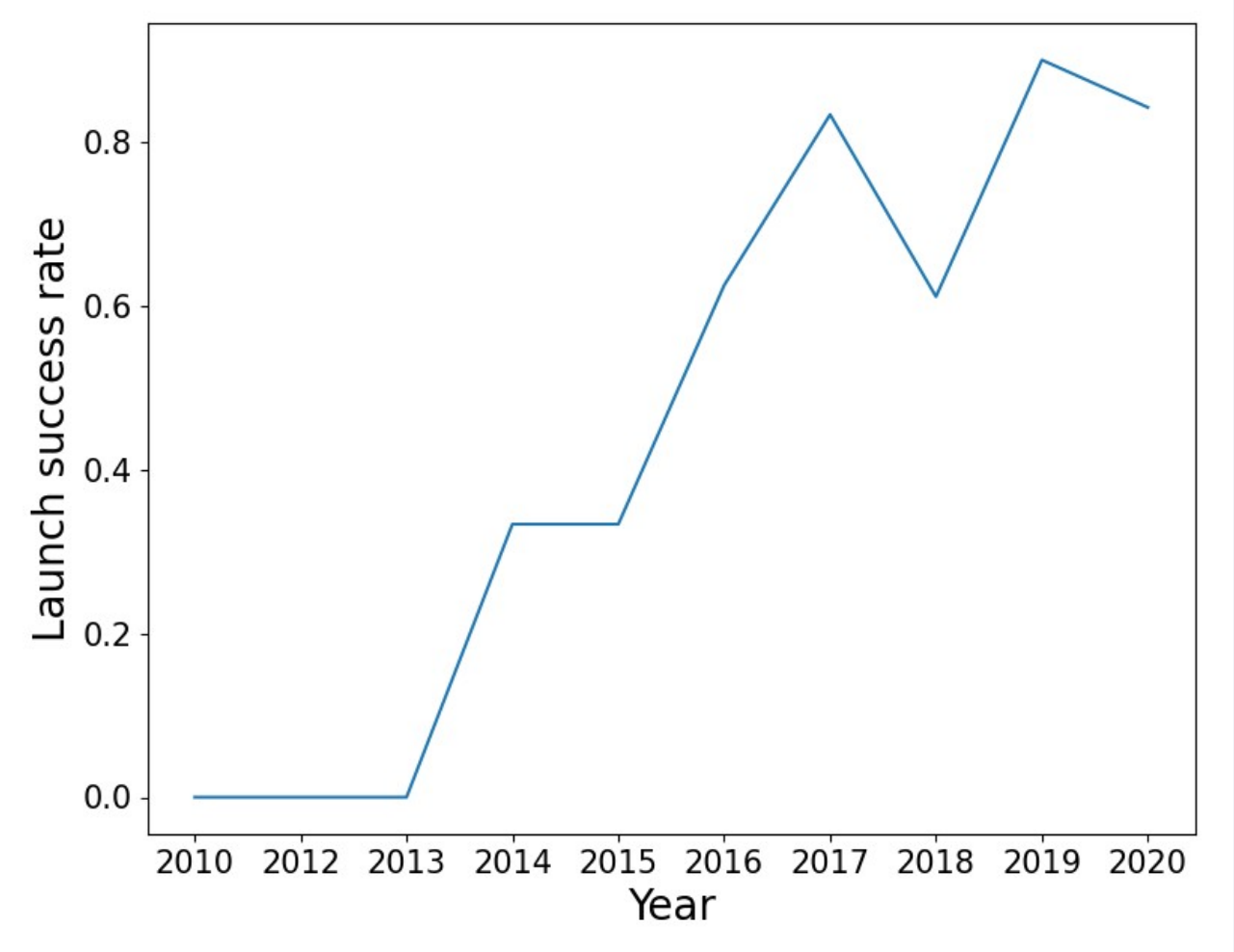
Payload vs. Orbit Type

- The highest payloads used orbit type VLEO.
- The lowest payloads used orbit types LEO, ISS, PO, ES-L1, SSO and HEO.
- The GTO orbit type was associates with the greatest number of launches for payload masses between 2000 kg and 7000 kg.



Launch Success Yearly Trend

- From 2010 to 2013, all launches failed.
- From 2013 onwards, the success rate increased from 0 % to 80 % in 2017.
- This declined to 60 % in 2018 but recovered afterwards.
- The most successful year was 2019 with 85 % of launches being successful.



All Launch Site Names

```
%%sql SELECT DISTINCT Launch_Site
FROM SPACEXTBL
```

- This query filtered out the distinct launch sites.

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

Launch Site Names Begin with 'CCA'

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

- Only launch sites with ‘CCA’ appearing in the name were selected, and limited to 5 entries.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Missio
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	
08-12-2010	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	
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	

Total Payload Mass

```
%%sql SELECT
SUM(PAYLOAD_MASS__KG_) AS
Total_Payload_Mass

FROM SPACEXTBL

WHERE Customer = 'NASA (CRS)'
```

Total_Payload_Mass
45596

- The payload masses of all launches were added together for the customer NASA.

Average Payload Mass by F9 v1.1

```
%%sql SELECT AVG(PAYLOAD_MASS__KG_) AS Total_Payload_Mass

FROM SPACEXTBL

WHERE Booster_Version LIKE '%F9 v1.1%'
```

Total_Payload_Mass
2534.6666666666665

- For the booster F9 v1.1, the average payload was calculated across all launches.

First Successful Ground Landing Date

```
%%sql SELECT MIN(Date)
FROM SPACEXTBL
WHERE "Landing _Outcome" = 'Success'
```

- The minimum (first) date was selected for launches marked as ‘successful’

MIN(Date)
02-03-2019

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql SELECT Booster_Version, "Landing
_Outcome"

FROM SPACEXTBL

WHERE PAYLOAD_MASS__KG_ < 6000
AND PAYLOAD_MASS__KG_ > 4000 AND
"Landing _Outcome" IS "Success (drone
ship)"
```

Booster_Version	Landing _Outcome
F9 FT B1022	Success (drone ship)
F9 FT B1026	Success (drone ship)
F9 FT B1021.2	Success (drone ship)
F9 FT B1031.2	Success (drone ship)

- For payloads between 4000-6000 kg, the booster version and landing outcome for successful launches were retrieved.

Total Number of Successful and Failure Mission Outcomes

```
%%sql SELECT COUNT
(Mission_Outcome) AS
Total_Mission_Outcomes FROM
SPACEXTBL
```

Total_Mission_Outcomes
101

- The number of successful and failed missions was found by counting the total number of missions.

Boosters Carried Maximum Payload

```
%%sql SELECT Booster_Version
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ IS
(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

- Booster versions are displayed where the payload mass carried was the maximum.

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

2015 Launch Records

```
%%sql SELECT SUBSTR(DATE,4,2), "Landing _Outcome", Booster_Version, Launch_Site

FROM SPACEXTBL

WHERE "Landing _Outcome" IS "Failure (drone ship)"

AND SUBSTR(DATE,7,4)='2015'
```

- Landing outcome, booster and launch site was retrieved for all failed launches in 2015.

SUBSTR(DATE,4,2)	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

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

```
%%sql SELECT Date, "Landing _Outcome", COUNT("Landing _Outcome")
FROM SPACEXTBL

WHERE "Landing _Outcome" LIKE '%Success%'

AND Date > '04-06-2010' AND Date < '20-03-2017'

GROUP BY "Landing _Outcome"
```

- Number of launches for each type of landing outcome were counted between 04/06/2010 and 20/03/2017.

Date	Landing _Outcome	COUNT("Landing _Outcome")
07-08-2018	Success	20
08-04-2016	Success (drone ship)	8
18-07-2016	Success (ground pad)	6

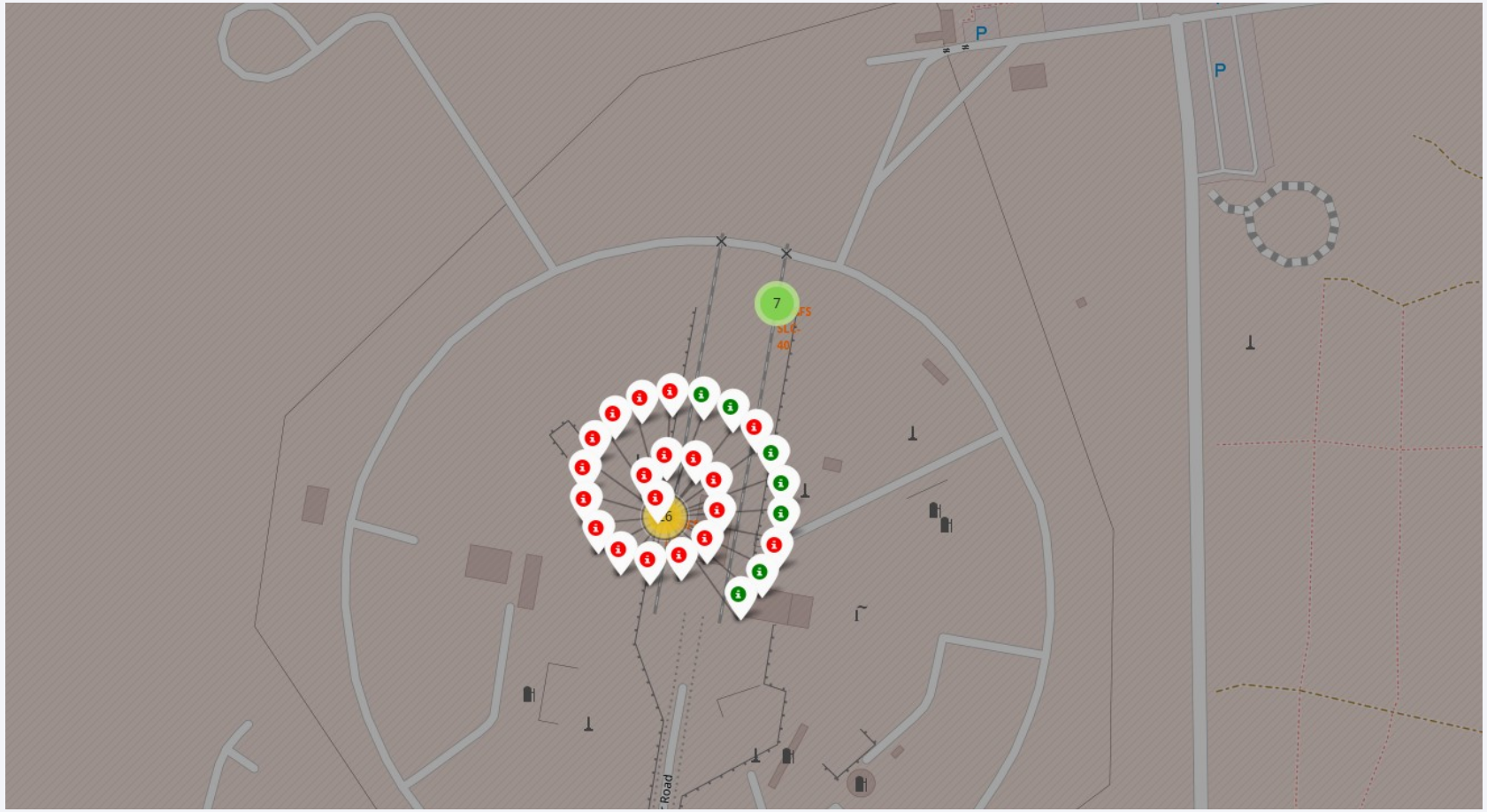
Section 3

Launch Sites Proximities Analysis

All launch sites in USA

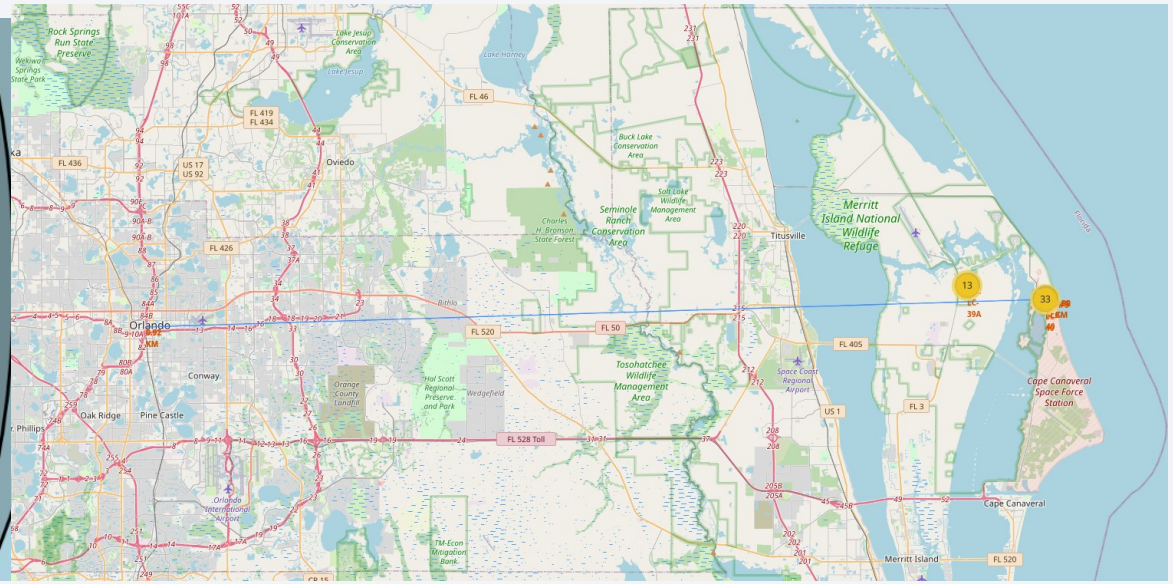
- Launch sites are located on the east coast and west coast of the USA close to the equator line.

Colour coded landing outcome of each launch



- Failed launches are marked in red, successful in green.

Launch site proximity to major infrastructure



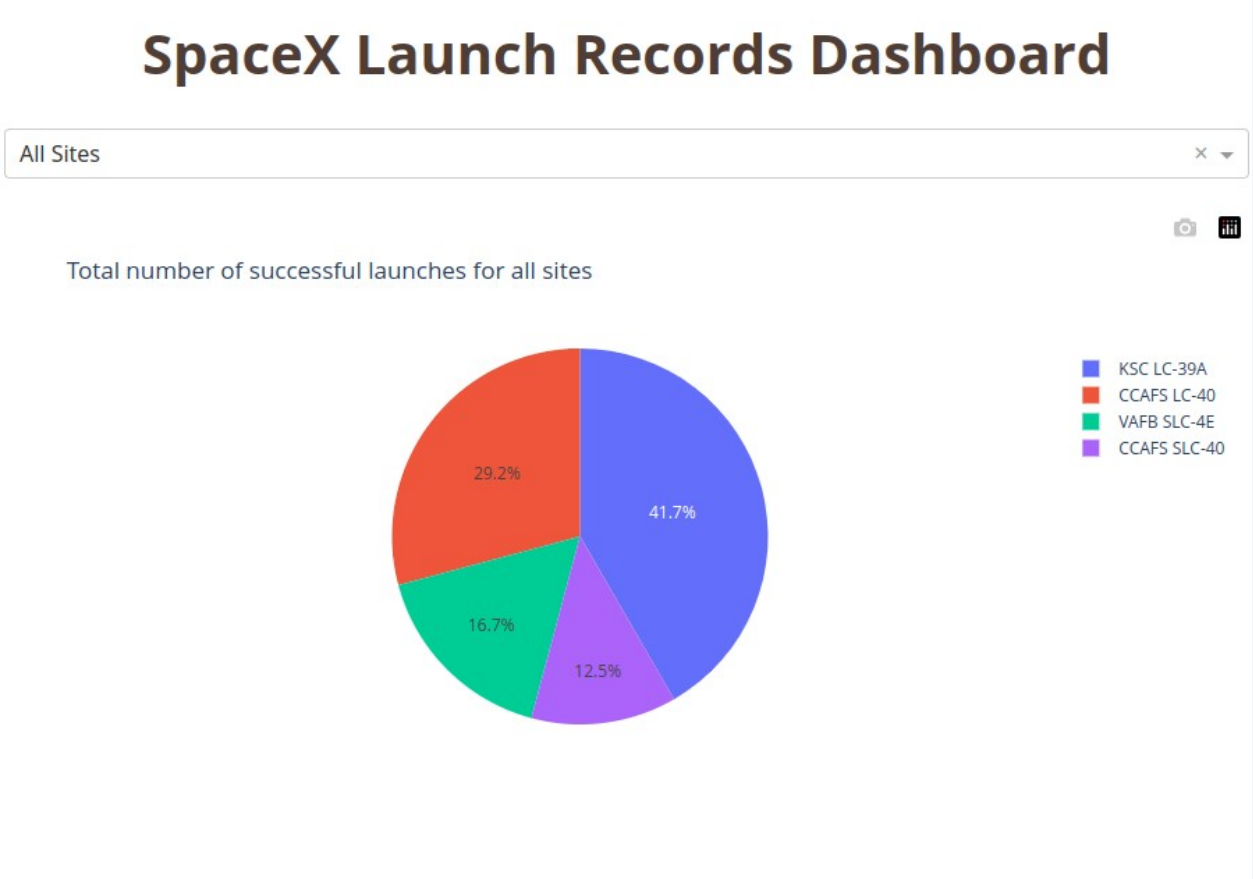
- Left: distance to coastline from east coast launch sites is 0.92 km.

Right: distance to Orlando from east coast launch sites is approximately 75 km.

Section 4

Build a Dashboard with Plotly Dash

Success rates of all launch sites

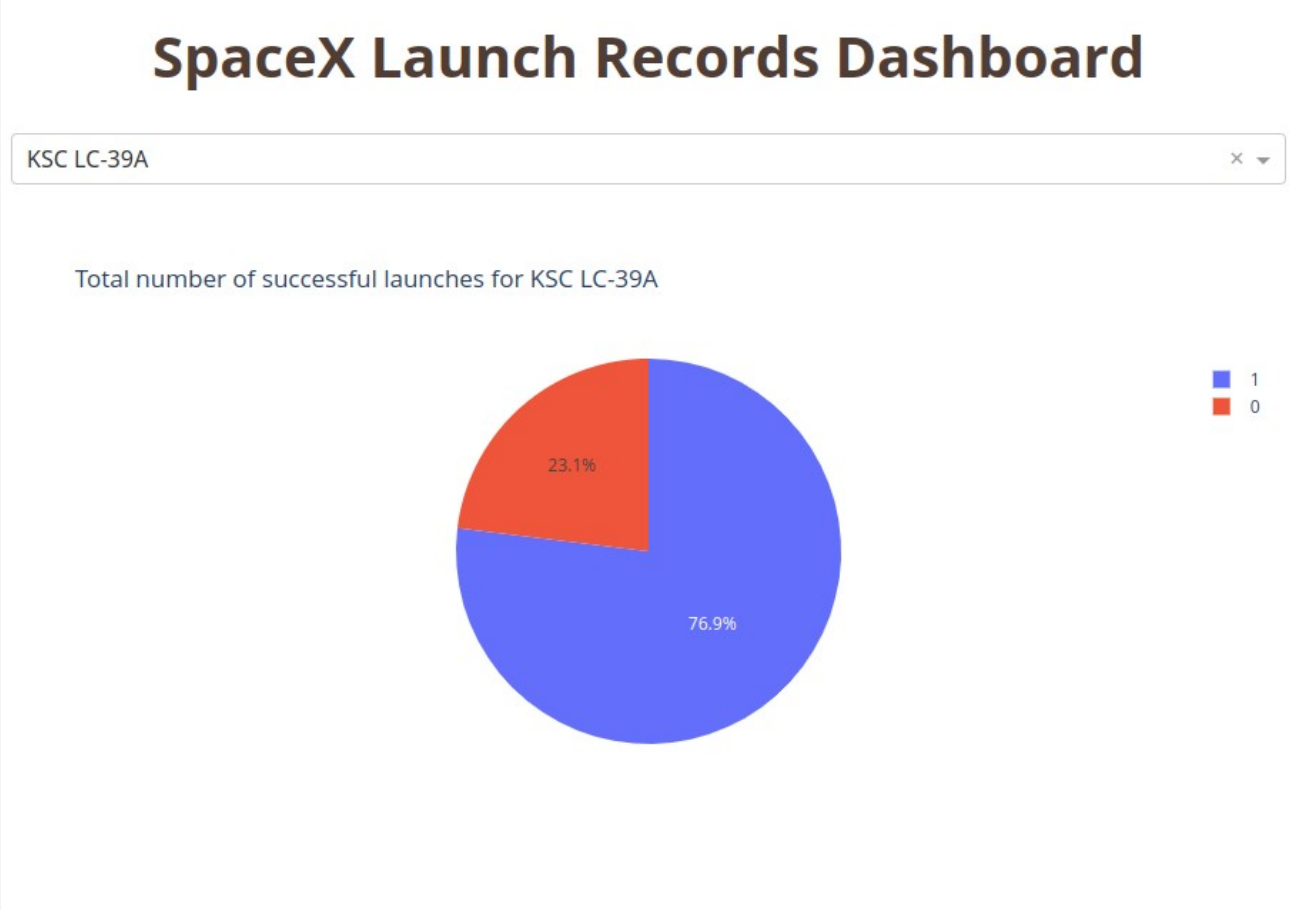


- A breakdown of launch success rates for each launch site.

The site with the greatest number of successful launches is KSC LC-39A

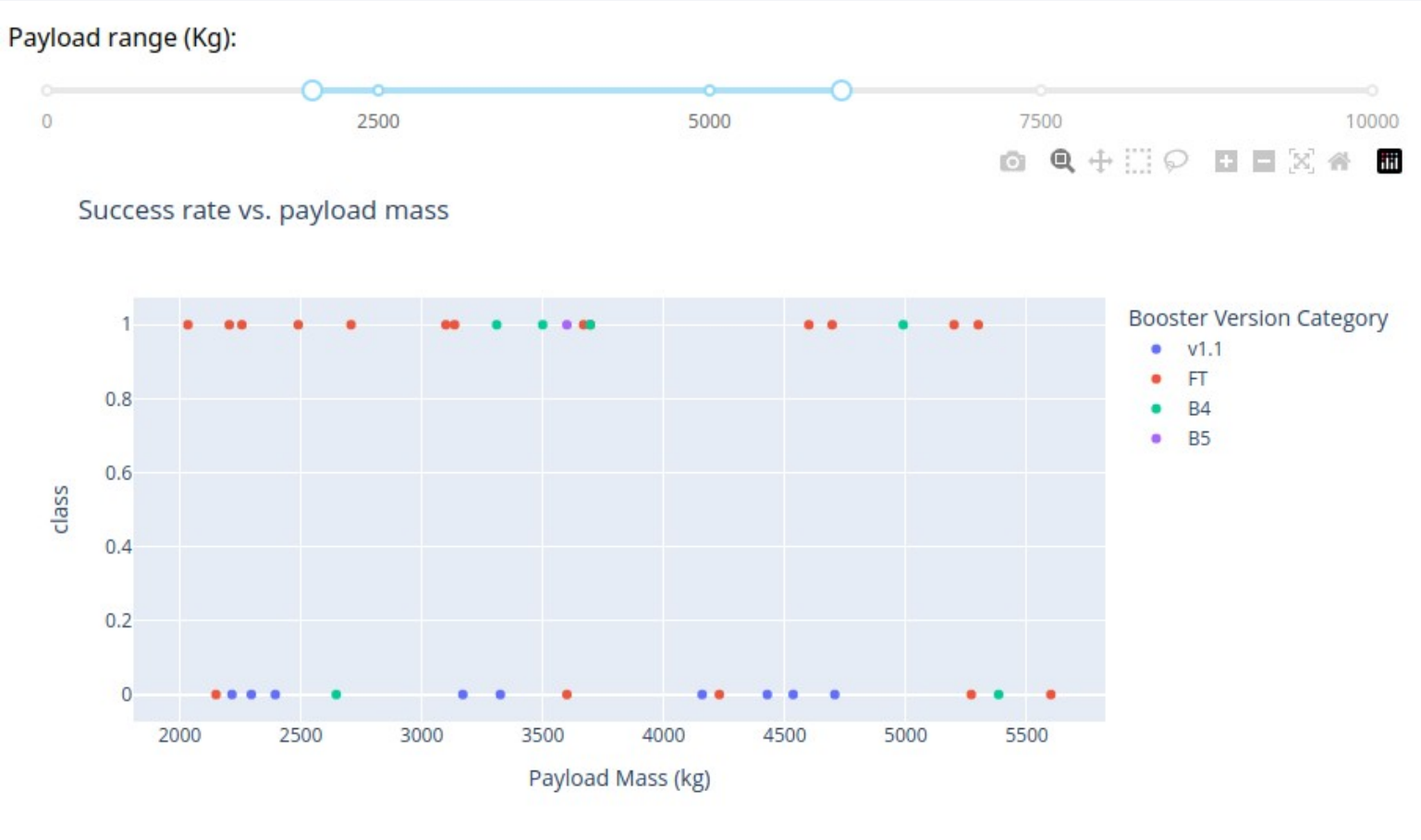
The site with greatest number of failures is CCAFS SLC-40.

Launch outcomes for launch site KSC LC-39A



- From the pie chart, we see that approximately three quarters of all launches were successful.

Success rate as function of payload mass



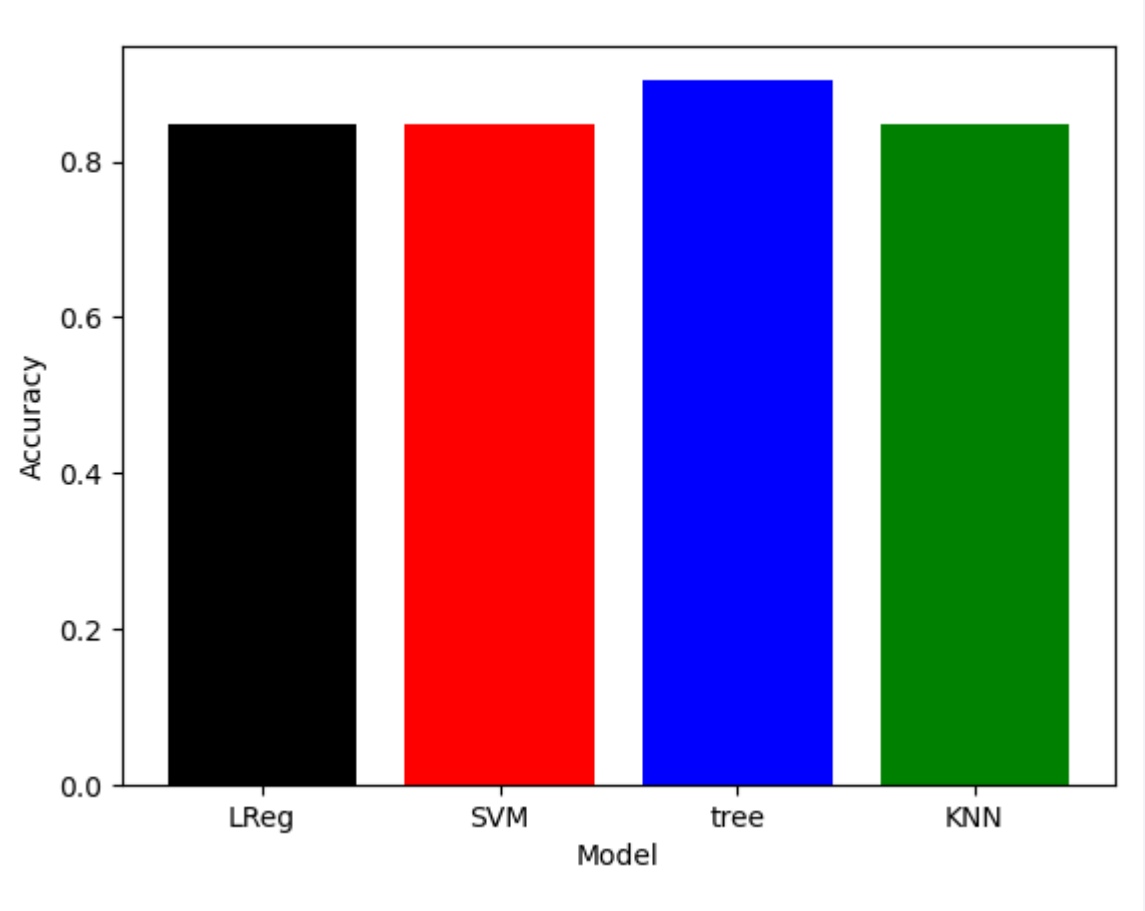
- Most successful launches occurred for payload masses between 2000 kg and 5250 kg.

Section 5

Predictive Analysis (Classification)

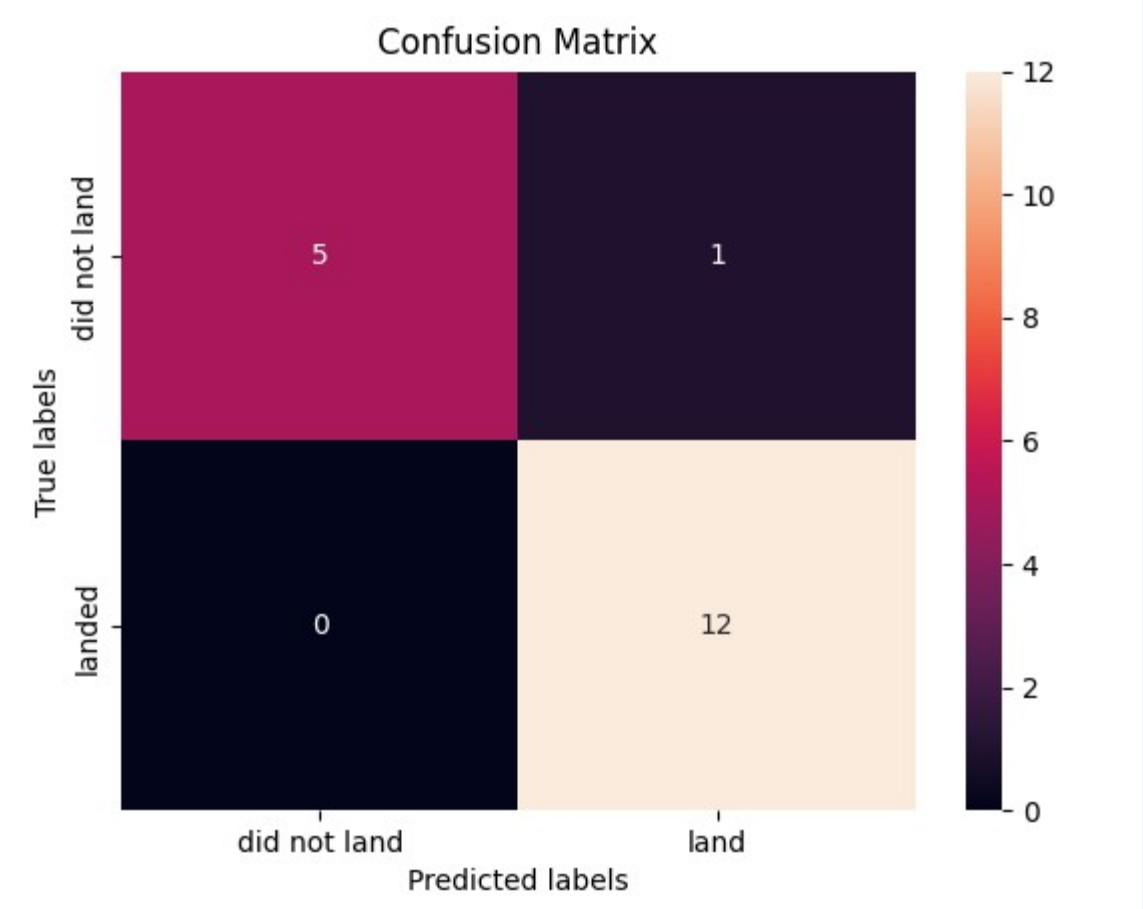
Classification Accuracy

- The tree model had the highest rate of accuracy at approximately 95 %.
- All other classification models had an accuracy of approximately 85 %.
- Therefore, the tree classification model was selected as the most suitable choice.



Confusion Matrix of the tree model

- The accuracy (correctness of all predictions) is 94.4 %.
- The precision (reliability of a positive prediction) is 92.3 %.
- The sensitivity (true positive rate) is 100 %.
- The likelihood of a false negative is zero.



Conclusions

- The orbit types ES-L1, GEO, HEO and SSO have a 100 % success rate, whereas SO has a 0 % success rate.

- The launch site with the highest success rate is KSC LC-39A, and the site with greatest number of failures is CCAFS SLC-40.

- Most successful launches occurred for payload masses between 2000 kg and 5250 kg.

- A decision tree model was created to predict launch success with an accuracy of 94.4 %.

Insights and outlook

- Further analysis could be conducted to evaluate the cost of various logistic options, i.e. whether train transport or highway transport of rocket boosters is more economical.

- Inquiries could be performed into improving the success of SO orbit type launches as all launches with this orbit type currently fail.

- Launch site CCAFS SLC-40 could be investigated as only 12.5 % of launches are successful from this site.

Thank you!