IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Dean Kennedy

12/07/2023

Outline

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

Executive Summary

This report aims to identify, through data science, the key variables to ensure a successful rocket landing – a vital process to reduce costs. The following methodology was used:

- **Collect** SpaceX data through various means
- **Wrangle** the data to determine a success
- **Explore** the data using typical visualisation techniques
- **Analyse** to produce necessary launch statistics and site information
- **Predict** landing outcomes based on previous launches

The key results of this report are as follows:

- Landing success has improved over time, and specific orbits (GEO, HEO, SSO and ES-L1) have high success rates.
- All predictive models produced similar results on the small dataset.

Introduction

SpaceX are a space tourism company which are attempting to reduce the cost of trips through reusable rockets. In this report, we attempt to use SpaceX data and predictive modelling to determine whether a particular launch will succeed; that is, the reusable rocket lands and can be reused.

The results of this report will be produced by analysing how factors such as payload mass, launch site and orbit can affect the rate of successful landings. Predictive models, specifically binary classification models, will be built with these factors to determine a landing success with a specific set of factors.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- 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 – SpaceX API

- Request data from the SpaceX API
- Use .json() to decode and using .json_normalize()
- Create a dataframe of the data by creating a dictionary of the data
- Ensure that the data is filtered to just include Falcon 9 launches
- Clean the data by replacing any missing values (if numerical use .mean())
- Create a csv file by exporting the dataframe

Data Collection - Scraping

- Scrape the data from Wikipedia
- Use BeautifulSoup to extract column names from the HTML table header
- Parse the HTML tables to collect data
- Create a dataframe using a dictionary of the data
- Export to a csv file

Data Wrangling

- Explore the data to determine the data labels
  - Unsuccessful
  - True Ocean
  - False Ocean
  - True RTLS
  - False RTLS
  - True ASDS
  - False ASDS
- Convert outcomes to binary form (one-hot encoding)
- Export the data to a csv file

EDA with Data Visualization

- Produce visualisations of the following:
    - Flight Number vs. Launch Site, Flight Number vs. Payload, Payload Mass vs. Launch Site and Payload Mass vs Orbit Type
- Analyse the data to determine the relationships between the above variables using appropriate visualisations

EDA with SQL

- We will look to display:
  - Names of unique launch
  - Total payload mass carried by NASA
  - Average payload mass carried by the booster F9 v1.1
- We will list:
  - The first successful landing on a ground pad
  - Which boosters had a successful landing on a drone ship with a specific mass
  - Successful and failed missions
  - Which launches failed to land on a drone ship
  - The total outcomes between specific dates

Build an Interactive Map with Folium

- Display launch sites and coordinates:
    - Show NASA with a blue circle and popup label
    - Show all launch sites using red circles with popup labels
- Display launch outcomes
    - Add green markers for successful
    - Add red markers for unsuccessful
    - Add to launch sites depending on landing rate success
- Display distance to coastlines
    - Show how close to specific geographical features a launch site is

Build a Dashboard with Plotly Dash

- Produce drop down lists of launch sites
- Display a pie chart of successful launches
- Display a slider of payload masses
- Display a scatter chart showing payload mass vs success rate

Predictive Analysis (Classification)

- Scale and split the data into test and train sets
- Create GridSearchCV objects for cv=10 parameter optimisation
- Apply GridSearchCV to different algorithms:
  - Logisitic regression
  - Support Vector Machines
  - Decision Trees
  - K-Nearest Neighbours
- Calculate accuracy (and other scores if required) for models and produce confusion matrices to identify the best model

Results

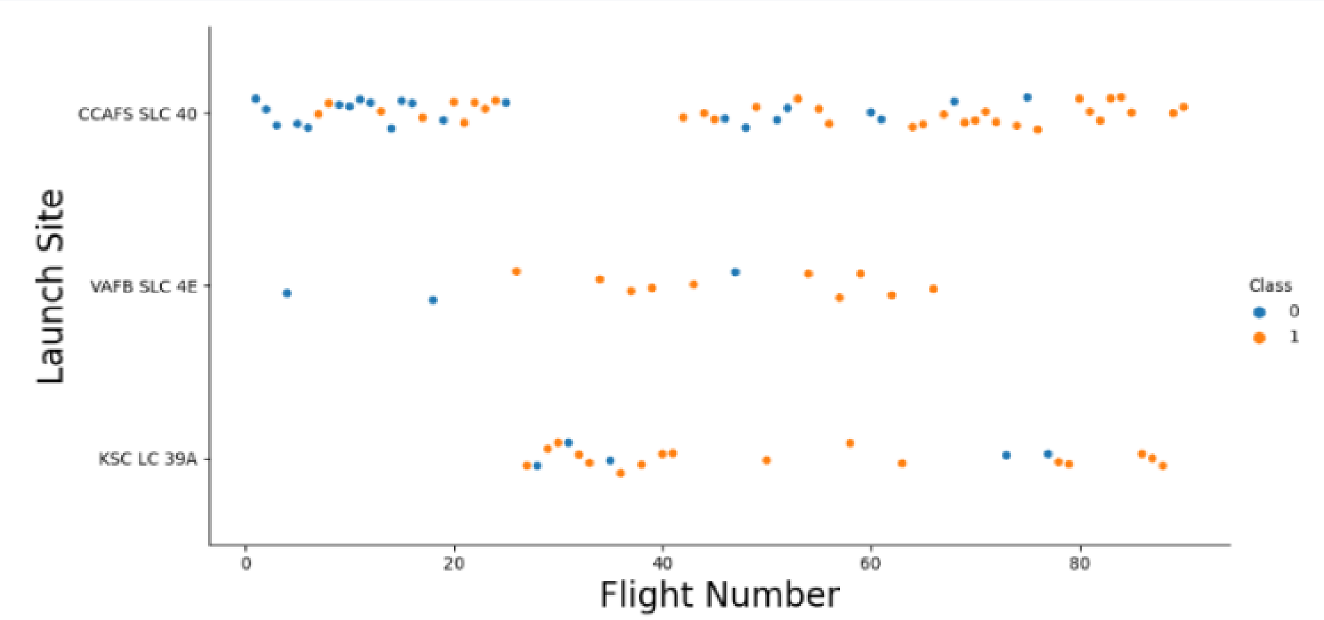
- Launch success has improved as the number of flights has increased
- The landing site KSC LC-39A has the best success rate
- The GEO, HEO, SSO and ES-L1 have 100% success rates
- All launch sites are in close proximity to the coastline
- The decision tree model had the best performance for predicting landing outcomes

Section 2

Insights drawn from EDA

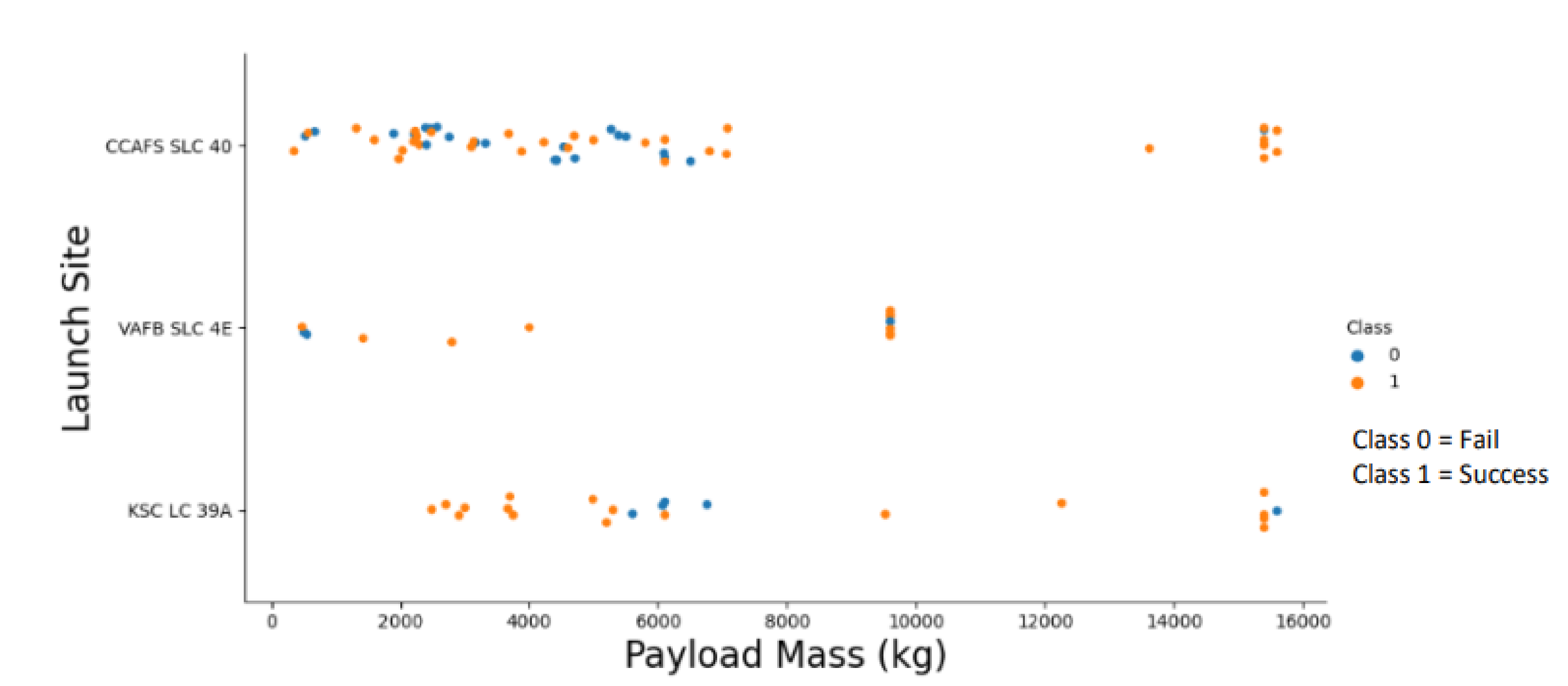
Flight Number vs. Launch Site

- The success rate improved as the number of flights increased. In the graph, blue is a fail and orange is a success.
- Most launches were from CCAFS SLC 40 but the other launch sites (VAFB SLC 4E and KSC LC 39A) have better success rates.



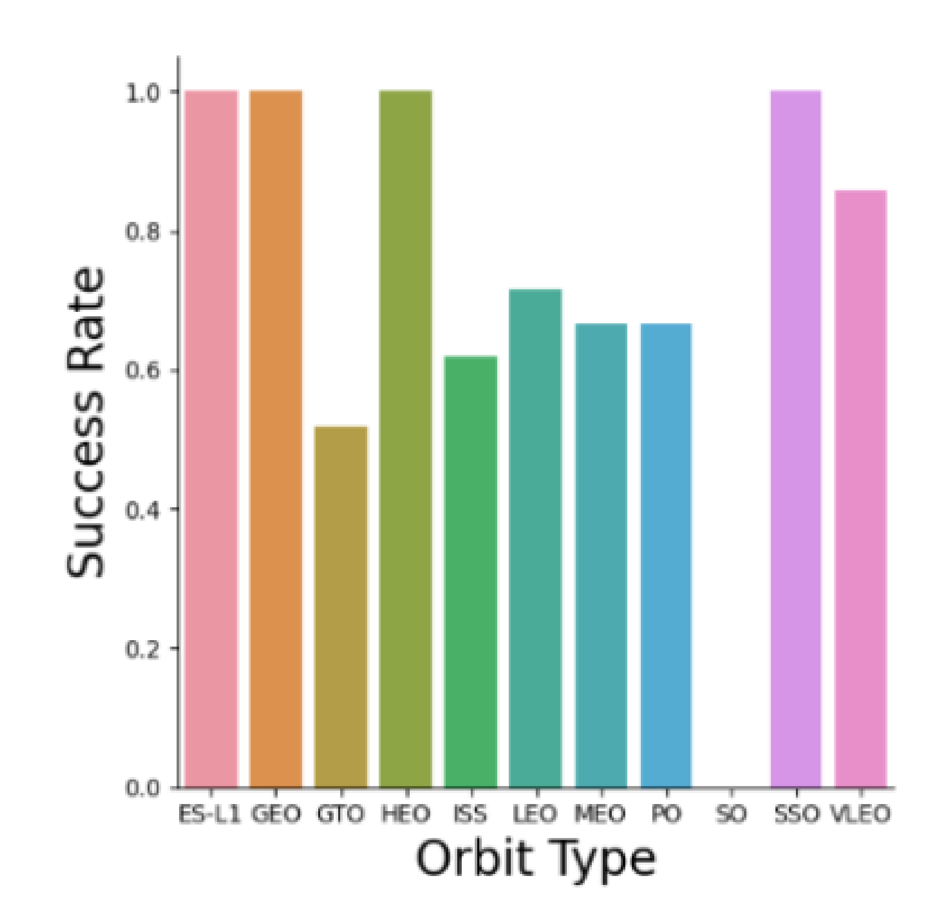
Payload vs. Launch Site

- Generally the launches with higher payload masses have a higher success rate as launches with a payload mass of over 7000kg have a very high success rate, though there is less data supporting this.



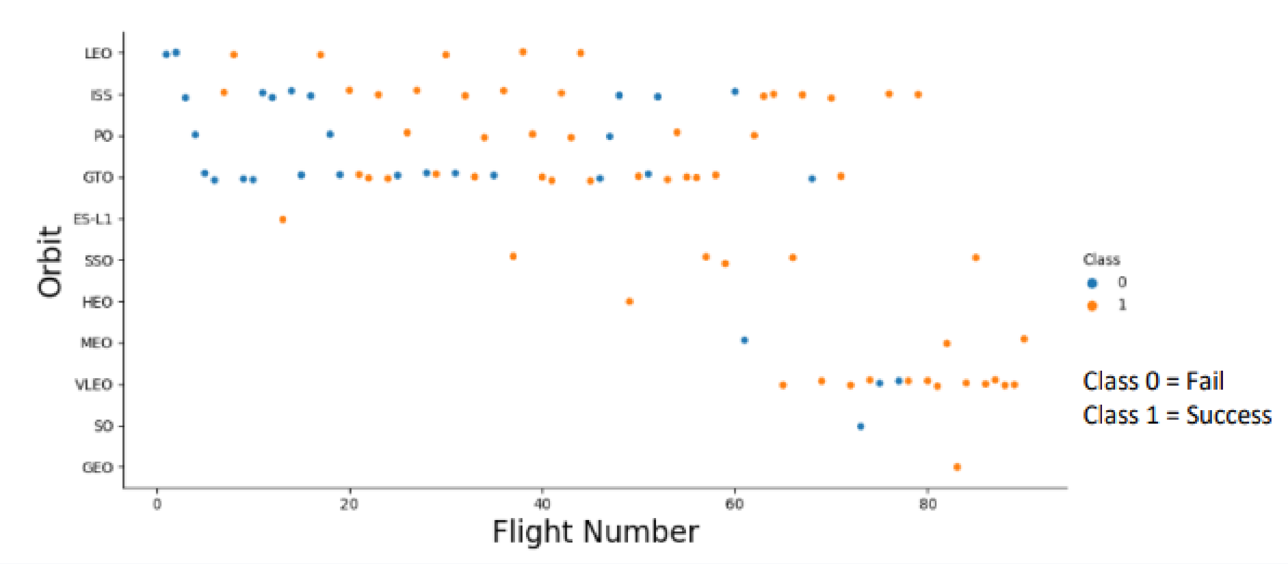
Success Rate vs. Orbit Type

- Four orbit types have a 100% success rate.
- One orbit has a 0% success rate.
- Six orbits have a success rate of between 50%-85%.



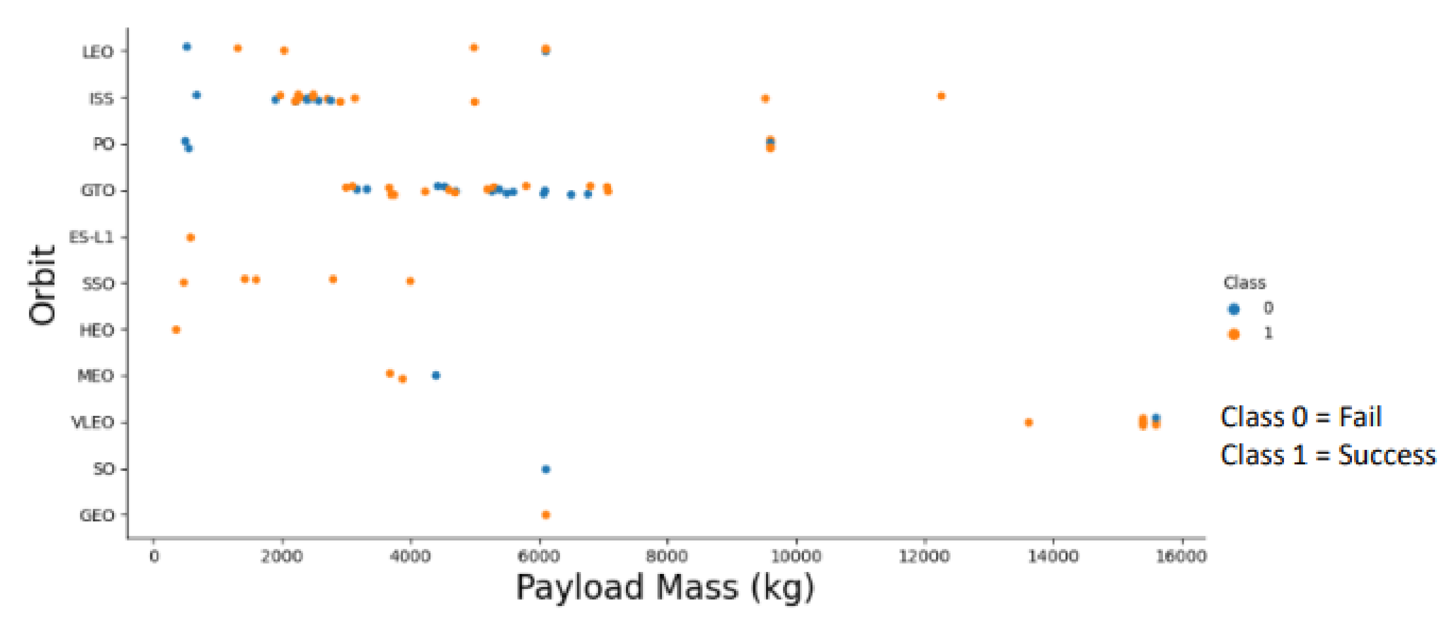
Flight Number vs. Orbit Type

- As the number of flights per orbit type increases, the success rate generally improves.
- GTO and ISS are the weakest examples of this trend but VLEO, LEO and PO are stronger examples of the trend.



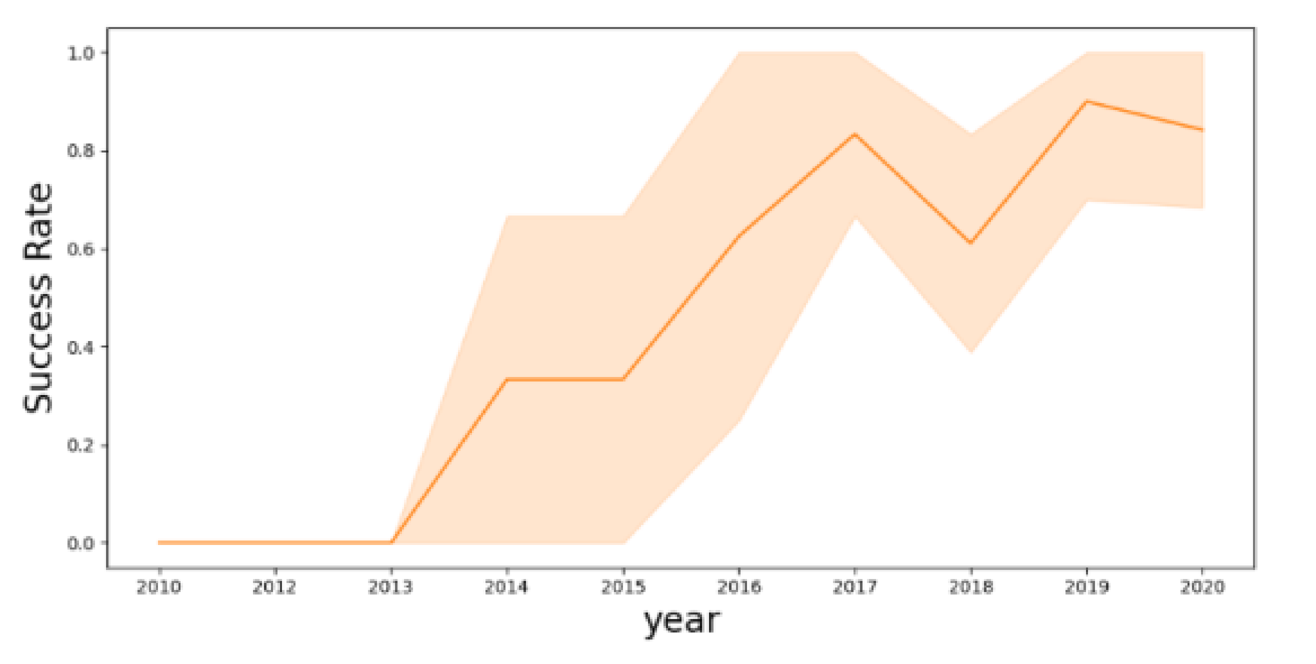
Payload vs. Orbit Type

- Generally, a heavy payload improves success rate but the GTO is an exception.



Launch Success Yearly Trend

- The success rate has generally improved over the years, but 2017-2018 and 2019-2020 are exceptions to this trend.



All Launch Site Names

- The launch sites have the following names:
  - CCASF LC-40
  - CCAFS SLC-40
  - KSC LC-39A
  - VAFB SLC-4E

```
[30]: %sql ibm_db_sa://yyy33800:dwNKg8J3L0IBd6CP@1bbf73c5
      %sql SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL;

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9
   sqlite:///my_data1.db
Done.
```

[30]:

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

- 5 examples of launches from CCAFS launch sites:

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

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
   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	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

- Total payload mass = 45,596kg by NASA launched boosters
- Average payload mass = 2,928kg by F9 boosters

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

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4l
   sqlite:///my_data1.db
Done.

    1
45596
```

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

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-
   sqlite:///my_data1.db
Done.

    1
2928
```

First Successful Ground Landing Date

- 12/22/2015

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

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b
   sqlite:///my_data1.db
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster mass between 4,000kg and 6,000kg:

```
%sql SELECT PAYLOAD \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;
```

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9
   sqlite:///my_data1.db
Done.
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

```
%sql SELECT BOOSTER_VERSION \
FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

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

- Failed landings in 2015 with all relevant details

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

month	Date	Booster_Version	Launch_Site	Landing _Outcome
01	10-01-2015	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	14-04-2015	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

```
%sql SELECT [Landing _Outcome], count(*) as count_outcomes \
FROM SPACEXTBL \
WHERE DATE between '04-06-2010' and '20-03-2017' group by [Landing _Outcome] order by count_outcomes DESC;
```

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

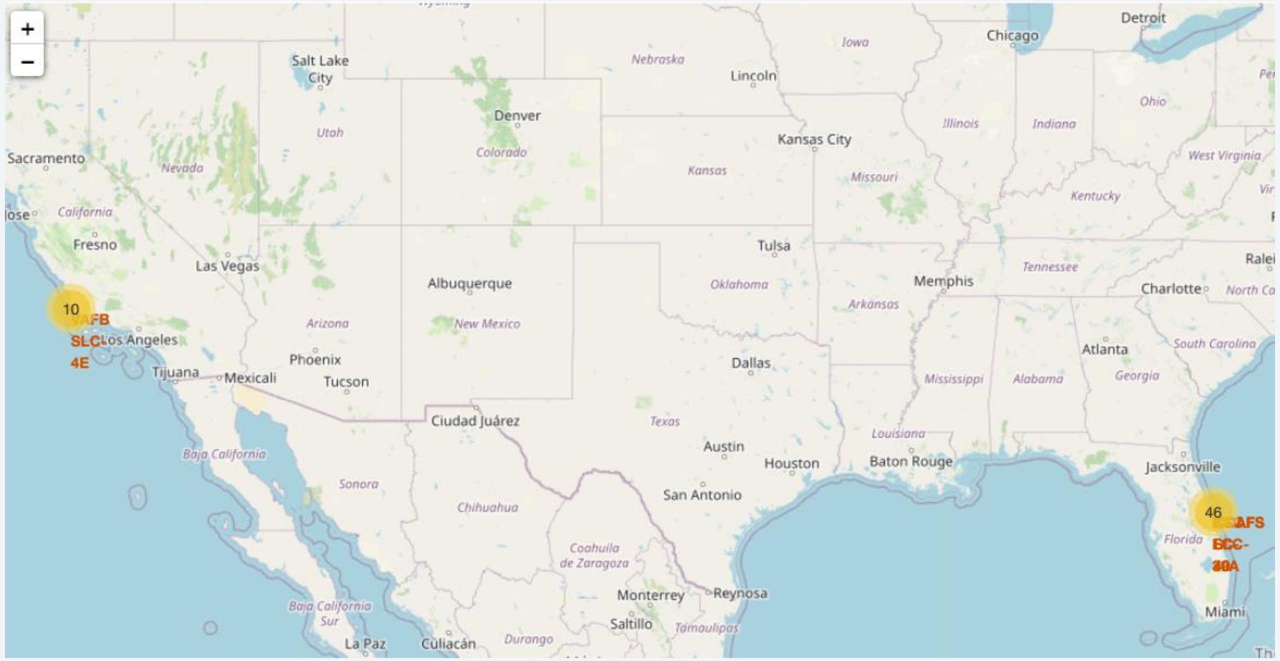
Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

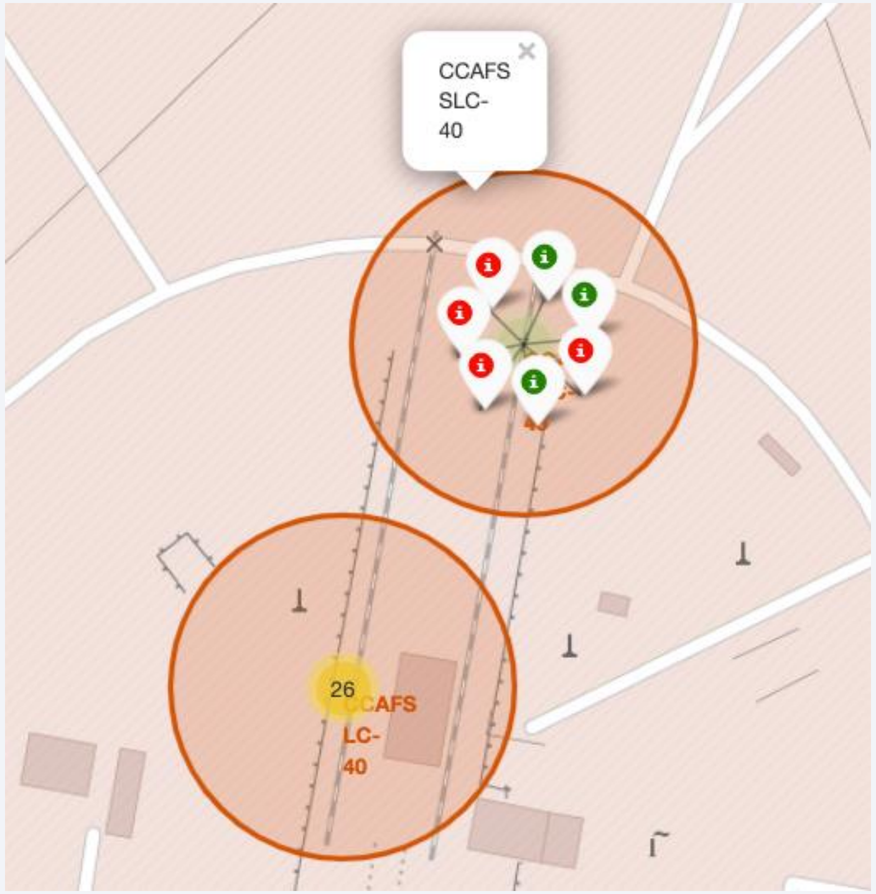
Section 3

Launch Sites Proximities Analysis

Launch Sites

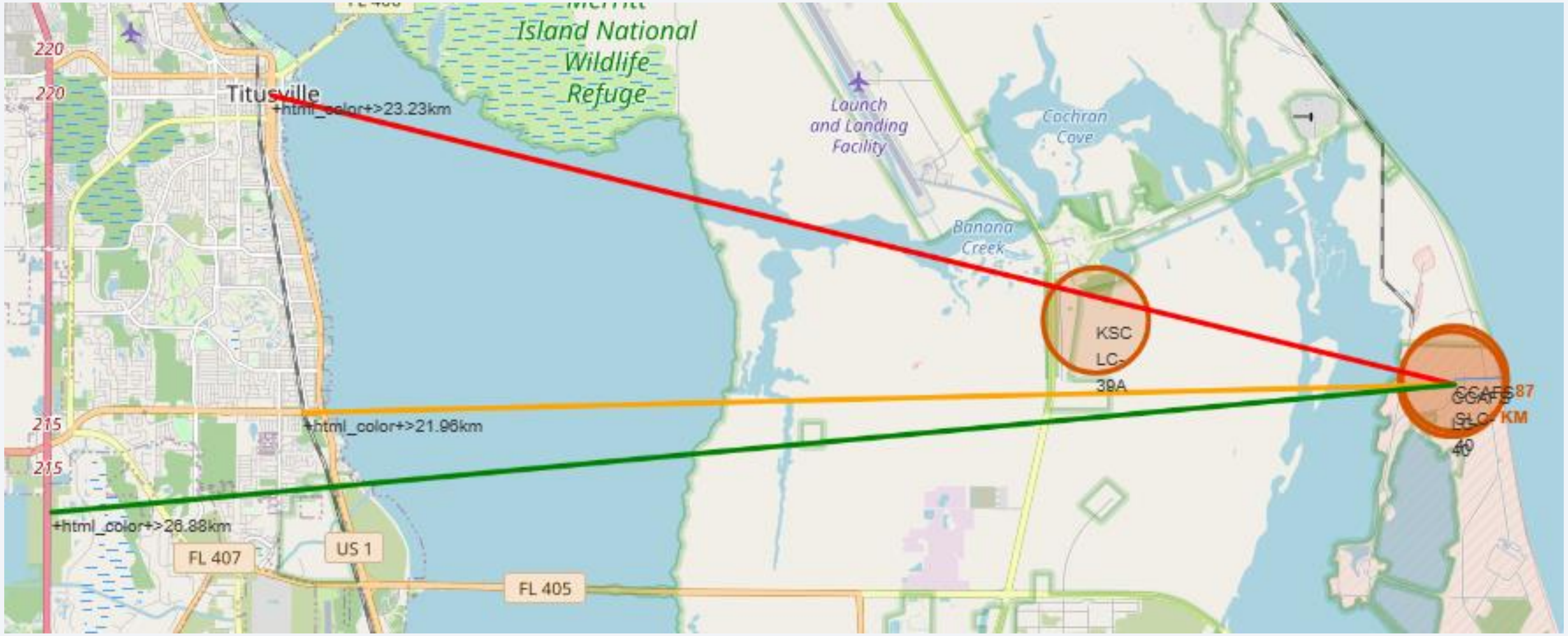
Launch Outcomes for Sites





Distance to Obstacles

- 0.86km to coastline, 21.96km to railway, 23.23km to city, 26.88km to motorway

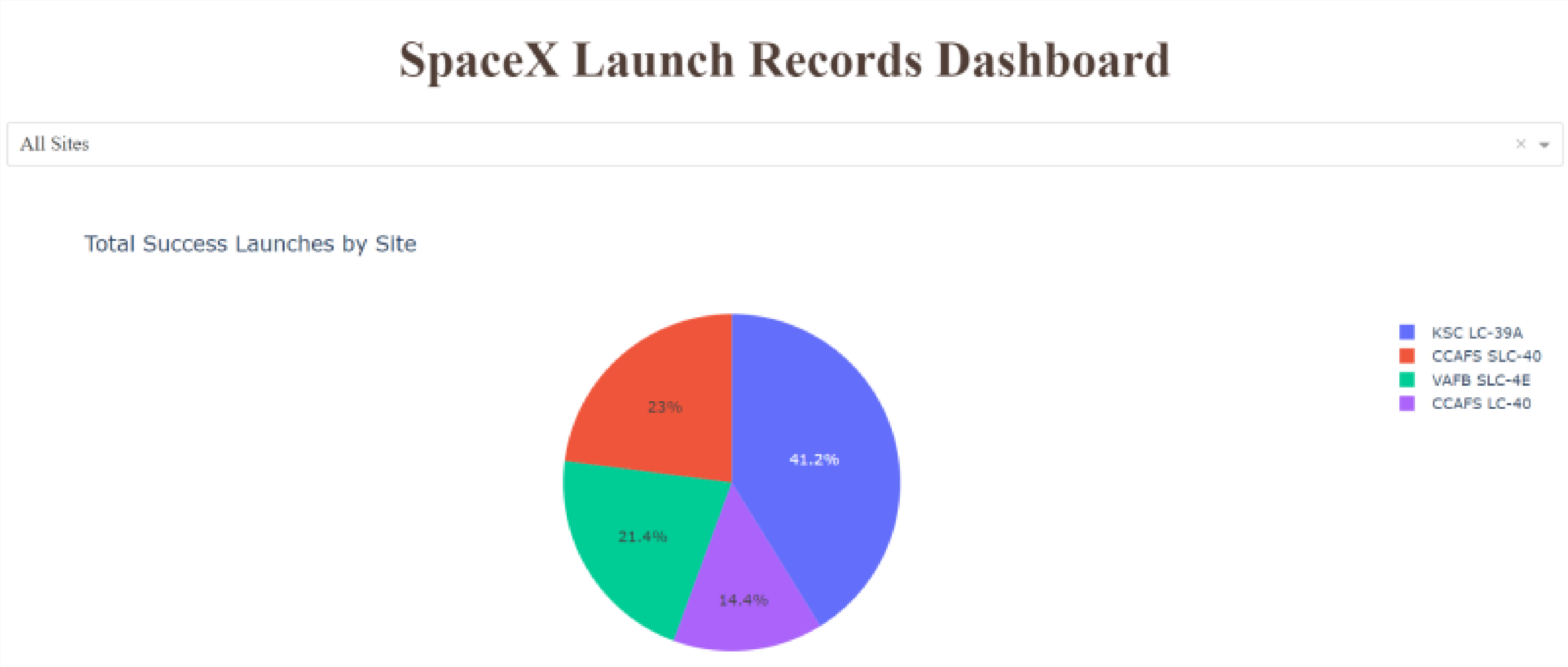


Section 4

Build a Dashboard with Plotly Dash

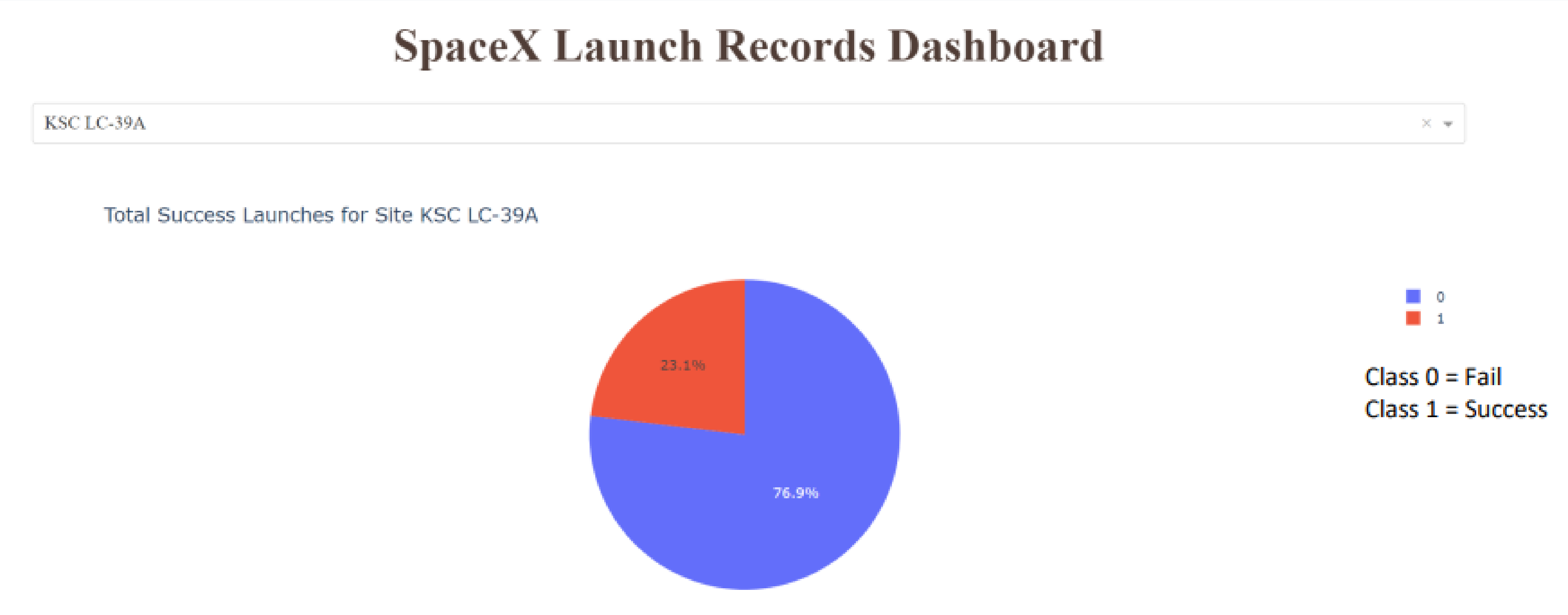
Launch Site Success Rate

- Most successful launches by KSC LC-39A at 41.2%



Launch Success Rate of KSC LC-39A

- 76.9% success rate



Payload Mass Success Rate

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- The classification models were all similar in their predictions.

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

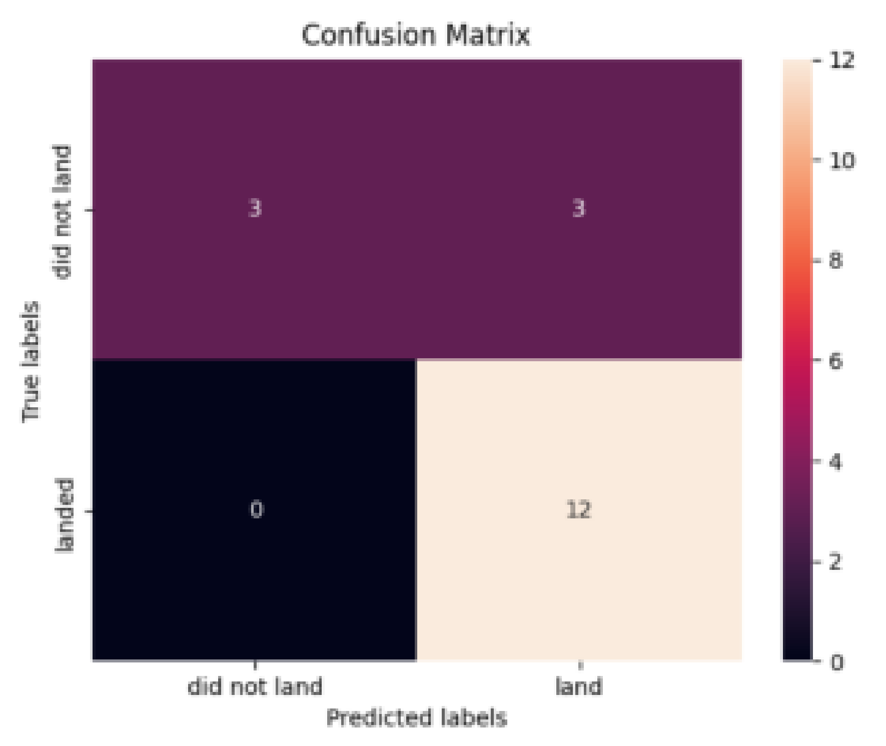
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.9017857142857142
Best params is : {'criterion': 'gini', 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}
```

Confusion Matrix

- All model confusion matrices were identical.
- All models produced 3 false positives.
- Accuracy = 83.3%
- F1 Score = 0.89



Conclusions

- Launch success rate improves over time
- Higher payload mass indicated a higher success rate
- Orbits GEO, HEO, SSO and ES-L1 have high success rates
- Launch sites are close to coasts and are as close to the equator as possible
- Larger dataset needed to improve predictive performance and understanding of features

Thank you!