IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

IBM Data Science Capstone Project
Conrad Kleykamp
7/24/24

OUTLINE

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

EXECUTIVE SUMMARY

Methodologies

- Data Collection
  - API, Web Scraping (BeautifulSoup)
- Data Wrangling
  - Missing values
  - Training label → 'Outcome' column
  - One Hot Encoding
- Exploratory Data Analysis (EDA)
  - SQL Queries
  - Visualization w/ Matplotlib, Seaborn
- Launch Sites Locations Analysis
  - Folium
- Launch Records Dashboard
  - Plotly Dash
- Predictive Analysis
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors Classifier

EXECUTIVE SUMMARY

General Results

- Binary Classification
  - ‘Outcome’ → 1 (Successful Landing)
  - ‘Outcome’ → 0 (Unsuccessful Landing)
- ~ 83.33% test set accuracy across all predictive methods
- ~ 88.88% train set accuracy for the Decision Tree Classifier

INTRODUCTION

Background

- Completed as part of IBM's Applied Data Science Capstone course
- SpaceY is a fictional commercial rocket launch provider that wishes to compete against SpaceX
- SpaceX's Falcon 9 rocket launches advertised to cost $62 million
    - Other rocket launch providers cost ~$165 million
    - SpaceX savings are due to the ability to land and reuse the first stage

INTRODUCTION

Business Problem

- How can a company like SpaceY bid against SpaceX for a rocket launch?
    - Must deduce the cost of a launch
- Using public SpaceX data, if we can determine whether the first stage will land, we can determine the cost of a launch
- This project will leverage multiple machine learning models to accurately predict the likelihood of a successful landing of the first stage rocket
- Important variables to consider:
    - Payload mass, launch site, orbit, number of flights

Section 1

Methodology

METHODOLOGY

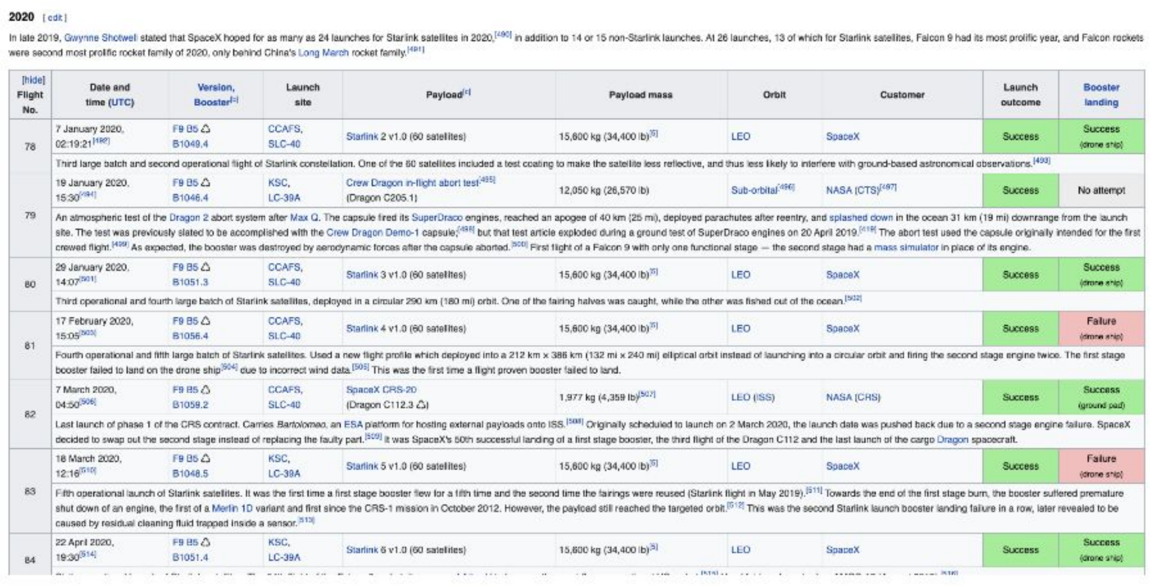
Overview

- 1: Data Collection
  - API, Web Scraping (BeautifulSoup)
- 2: Data Wrangling
  - Missing values
  - Training label → ‘Outcome’ column
  - One Hot Encoding
- 3: Exploratory Data Analysis (EDA)
  - SQL Queries
  - Visualization w/ Matplotlib, Seaborn
- 4: Launch Sites Locations Analysis
  - Folium
- 5: Launch Records Dashboard
  - Plotly Dash
- 6: Predictive Analysis
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree Classifier
  - K Nearest Neighbors Classifier

DATA COLLECTION

Overview

- SpaceX API
    - Historical launch data from SpaceX REST API
- Web Scraping
    - Historical launch data from Wikipedia page
    - [‘List of Falcon 9 and Falcon Heavy Launches’]



2020 [edit]

In late 2019, Gwynne Shotwell stated that SpaceX hoped for as many as 24 launches for Starlink satellites in 2020,[490] in addition to 14 or 15 non-Starlink launches. At 26 launches, 13 of which for Starlink satellites, Falcon 9 had its most prolific year, and Falcon rockets were second most prolific rocket family of 2020, only behind China's Long March rocket family.[491]

[hide] Flight No.	Date and time (UTC)	Version, Booster[c]	Launch site	Payload[c]	Payload mass	Orbit	Customer	Launch outcome	Booster landing
78	7 January 2020, 02:19:21[492]	F9 B5 ♺ B1049.4	CCAFS, SLC-40	Starlink 2 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third large batch and second operational flight of Starlink constellation. One of the 60 satellites included a test coating to make the satellite less reflective, and thus less likely to interfere with ground-based astronomical observations.[493]								
79	19 January 2020, 15:30[494]	F9 B5 ♺ B1046.4	KSC, LC-39A	Crew Dragon in-flight abort test[495] (Dragon C205.1)	12,050 kg (26,570 lb)	Sub-orbital[496]	NASA (CTS)[497]	Success	No attempt
	An atmospheric test of the Dragon 2 abort system after Max Q. The capsule fired its SuperDraco engines, reached an apogee of 40 km (25 mi), deployed parachutes after reentry, and splashed down in the ocean 31 km (19 mi) downrange from the launch site. The test was previously slated to be accomplished with the Crew Dragon Demo-1 capsule,[498] but that test article exploded during a ground test of SuperDraco engines on 20 April 2019.[419] The abort test used the capsule originally intended for the first crewed flight.[499] As expected, the booster was destroyed by aerodynamic forces after the capsule aborted.[500] First flight of a Falcon 9 with only one functional stage — the second stage had a mass simulator in place of its engine.								
80	29 January 2020, 14:07[501]	F9 B5 ♺ B1051.3	CCAFS, SLC-40	Starlink 3 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)
	Third operational and fourth large batch of Starlink satellites, deployed in a circular 290 km (180 mi) orbit. One of the fairing halves was caught, while the other was fished out of the ocean.[502]								
81	17 February 2020, 15:05[503]	F9 B5 ♺ B1056.4	CCAFS, SLC-40	Starlink 4 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fourth operational and fifth large batch of Starlink satellites. Used a new flight profile which deployed into a 212 km × 386 km (132 mi × 240 mi) elliptical orbit instead of launching into a circular orbit and firing the second stage engine twice. The first stage booster failed to land on the drone ship[504] due to incorrect wind data.[505] This was the first time a flight proven booster failed to land.								
82	7 March 2020, 04:50[506]	F9 B5 ♺ B1059.2	CCAFS, SLC-40	SpaceX CRS-20 (Dragon C112.3 ♺)	1,977 kg (4,359 lb)[507]	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
	Last launch of phase 1 of the CRS contract. Carries Bartolomeo, an ESA platform for hosting external payloads onto ISS.[508] Originally scheduled to launch on 2 March 2020, the launch date was pushed back due to a second stage engine failure. SpaceX decided to swap out the second stage instead of replacing the faulty part.[509] It was SpaceX's 50th successful landing of a first stage booster, the third flight of the Dragon C112 and the last launch of the cargo Dragon spacecraft.								
83	18 March 2020, 12:16[510]	F9 B5 ♺ B1048.5	KSC, LC-39A	Starlink 5 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Failure (drone ship)
	Fifth operational launch of Starlink satellites. It was the first time a first stage booster flew for a fifth time and the second time the fairings were reused (Starlink flight in May 2019).[511] Towards the end of the first stage burn, the booster suffered premature shut down of an engine, the first of a Merlin 1D variant and first since the CRS-1 mission in October 2012. However, the payload still reached the targeted orbit.[512] This was the second Starlink launch booster landing failure in a row, later revealed to be caused by residual cleaning fluid trapped inside a sensor.[513]								
84	22 April 2020, 19:30[514]	F9 B5 ♺ B1051.4	KSC, LC-39A	Starlink 6 v1.0 (60 satellites)	15,600 kg (34,400 lb)[5]	LEO	SpaceX	Success	Success (drone ship)

Example of launch records in HTML format prior to web scraping

DATA COLLECTION

SpaceX API

- Task 1:
    - Requested and parsed the SpaceX launch data using the GET request
- Task 2:
    - Filtered the dataframe to only include Falcon 9 launches
- Task 3:
    - Replaced null values in ‘PayloadMass’ column with mean ‘PayloadMass’ value

Github: 1-Data-Collection-API.ipynb

DATA COLLECTION

Web Scraping (Beautiful Soup)

- Task 1:
  - Requested the Falcon 9 Launch Wiki page from its URL
- Task 2:
  - Extracted all column/variable names from the HTML table header
- Task 3:
  - Created a dataframe by parsing the launch HTML tables

[Github: 2-Data-Collection-with-Web-Scraping.ipynb]

DATA WRANGLING

- Task 1:
  - Calculated the number of launches on each site
- Task 2:
  - Calculated the number and occurrence of each orbit
- Task 3:
  - Calculated the number and occurrence of mission outcome of the orbits
- Task 4:
  - Created a landing outcome label from ‘Outcome’ column
  - ‘Class’ → 1 if successful, 0 if unsuccessful

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

Landing Outcomes, n=90
~66% Success Rate

Github: 3-Data-Wrangling.ipynb

EDA

SQL

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

SQL Query #7

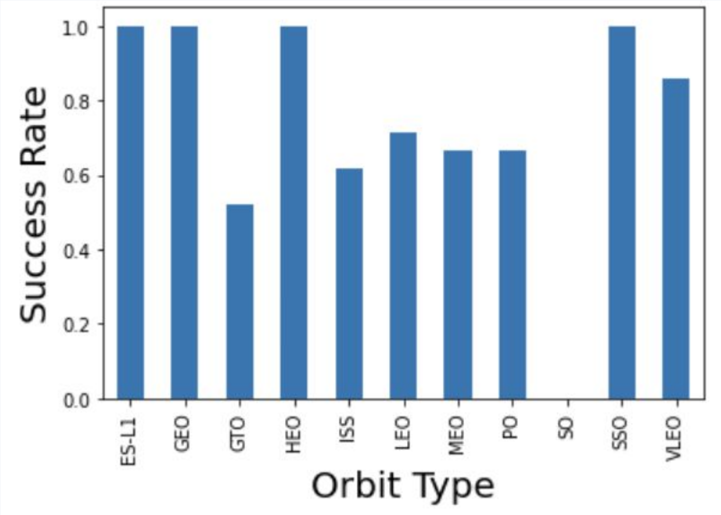
- Loaded data into IBM DB2 instance
- Completed queries:
  - 1: Displayed the names of unique launch sites
  - 2: Displayed 5 records where launch sites began with the string ‘CCA’
  - 3: Displayed the total payload mass carried by boosters launched by NASA (CRS)
  - 4: Displayed the average payload mass carried by booster version F9 v1.1
  - 5: Listed the date when the first successful landing outcome in ground pad was achieved
  - 6: Listed the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - 7: Listed the total number of successful and failure mission outcomes
  - 8: Listed the names of the booster versions which have carried the maximum payload mass
  - 9: Listed the failed landing outcomes in drone ship, along with their booster versions and launch site names for the year 2015
  - 10: Ranked the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order

Github: 4-EDA-with-SQL.ipynb

EDA

Data Visualization

- Scatter Point Charts (w/ landing outcome overlaid)
  - Flight Number x Launch Site
  - Payload x Launch Site
  - Flight Number x Orbit Type
  - Payload x Orbit Type
- Bar Chart
  - Orbit Type x Success Rate
- Line Chart
  - Year x Success Rate



Bar Chart: Orbit Type x Success Rate

[Github: 5-EDA-with-Visualization.ipynb]

LAUNCH SITES LOCATIONS ANALYSIS

Folium

- Task 1:
  - Marked all launch sites on a map
- Task 2:
  - Marked the successful/unsuccessful launches for each site on the map
- Task 3:
  - Calculated the distances between a launch site to its proximities

Interactive map showing successful/unsuccessful launches at the CCAFS SLC-40 launch site

[Github: 6-Launch-Sites-Locations-Analysis-with-Folium.ipynb]

LAUNCH RECORDS DASHBOARD

Plotly Dash

- Task 1:
  - Added a Launch Site drop-down input component
- Task 2:
  - Added a pie chart showing Success Rate based upon selected Launch Site
- Task 3:
  - Added a range slider to select Payload Mass
- Task 4:
  - Added a scatter chart showing Payload Mass x Landing Outcome

Github: 7-Interactive-Dashboard-with-Plotly-Dash

PREDICTIVE ANALYSIS

- Task 1:
  - Created a NumPy array from the ‘Class’ column and assigned it to the variable Y
- Task 2:
  - Standardized the data and assigned it to the variable X
- Task 3:
  - Split the X and Y data into training and test sets, with test size as 20%
- Task 4:
  - Fit the training data to the following models
    - Logistic Regression
    - Support Vector Machine
    - Decision Tree Classifier
    - K Nearest Neighbors Classifier
- Task 5:
  - Used GridSearchCV to identify best parameters for each model
- Task 6:
  - Calculated the accuracy on the test data using the optimal parameters for each model
- Task 7:
  - Plotted confusion matrices for each model’s predictions
- Task 8:
  - Compared the accuracies of each model

Github: 8-Machine-Learning-Prediction.ipynb

Section 2

Insights drawn from EDA

INSIGHTS: EDA - SQL

Launch Site Names

In [5]:

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

```
 * ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde0
0.databases.appdomain.cloud:31321/bludb
Done.
```

Out[5]:

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

INSIGHTS: EDA - SQL

Launch Site Names Begin with ‘CCA’

In [9]:

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

```
* ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[9]:

launch_site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

INSIGHTS: EDA - SQL

Total Payload Mass (NASA CRS Boosters)

In [10]:

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

```
 * ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[10]:

1
45596

INSIGHTS: EDA - SQL

Average Payload Mass by F9 v1.1 Booster

In [11]:

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTBL
WHERE Booster_Version LIKE 'F9 v1.0%';
```

```
 * ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde0
0.databases.appdomain.cloud:31321/bludb
Done.
```

Out[11]:

1
340

INSIGHTS: EDA - SQL

First Successful Ground Landing Date

```
In [30]:

%%sql
SELECT MIN(Date)
FROM SPACEXTBL
WHERE Landing__Outcome = 'Success (ground pad)';

 * ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde0
0.databases.appdomain.cloud:31321/bludb
Done.
Out[30]:
```

1
2015-12-22

INSIGHTS: EDA - SQL

Successful Drone Ship Landing with Payload Between 4000 and 6000 Kg

In [33]:

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

```
 * ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[33]:

booster_version
F9 FT B1021.1
F9 FT B1023.1
F9 FT B1029.2
F9 FT B1038.1
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

INSIGHTS: EDA - SQL

Total Number of Successful and Failure Mission Outcomes

In [44]:

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

```
* ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321/bludb
Done.
```

Out[44]:

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

INSIGHTS: EDA - SQL

Boosters Carrying Maximum Payload

In [57]:

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

```
* ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde0
0.databases.appdomain.cloud:31321/bludb
Done.
```

Out[57]:

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

INSIGHTS: EDA - SQL

2015 Launch Records

In [62]:

```
%%sql
SELECT LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE
FROM SPACEXTBL
WHERE Landing__Outcome = 'Failure (drone ship)'
    AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde0
0.databases.appdomain.cloud:31321/bludb
Done.
```

Out[62]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

INSIGHTS: EDA - SQL

Ranking Landing Outcomes (2010-06-04 – 2017-03-20)

In [70]:

```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS TOTAL_NUMBER
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME
ORDER BY TOTAL_NUMBER DESC
```

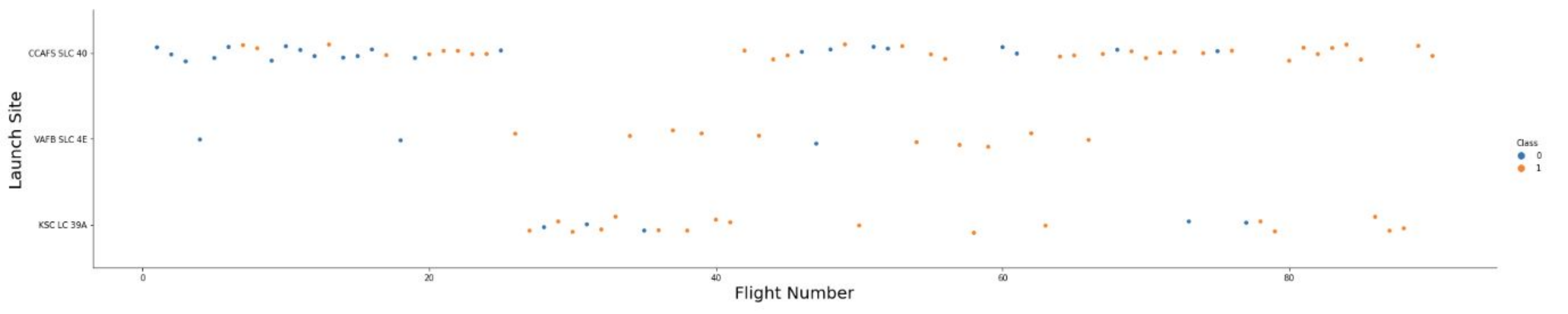
```
* ibm_db_sa://xcg80731:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.c1ogj3sd0tgtu0lqde0
0.databases.appdomain.cloud:31321/bludb
Done.
```

Out[70]:

landing__outcome	total_number
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

INSIGHTS: EDA - DATA VISUALIZATION

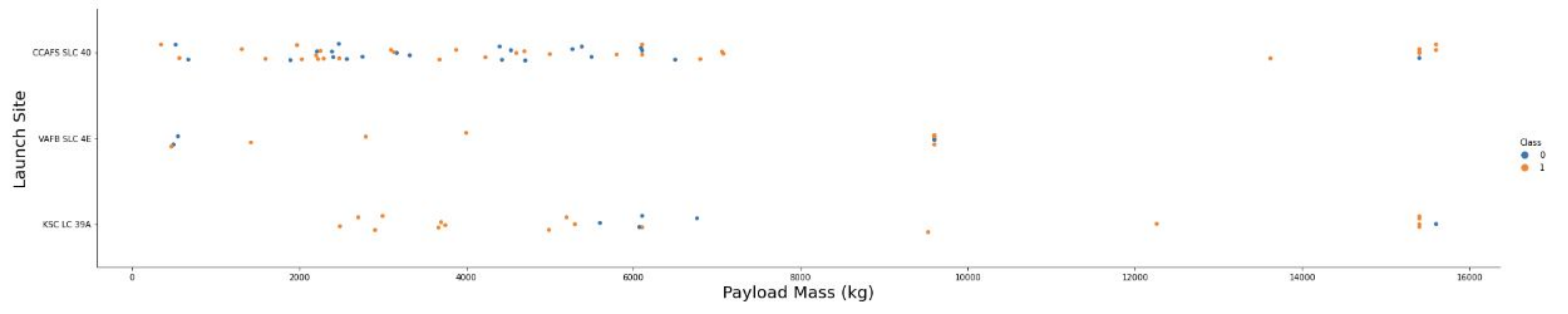
Flight Number x Launch Site



- The earliest flights tended to fail more often, while the latest flights tended to succeed
- The CCAFS SLC 40 launch site had the majority of launches

INSIGHTS: EDA - DATA VISUALIZATION

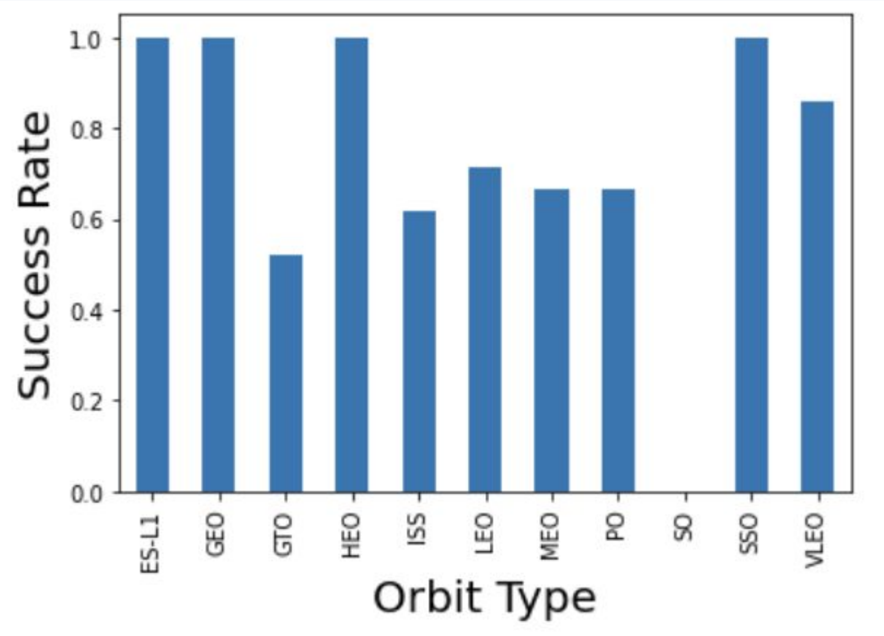
Payload x Launch Site



- Overall, a higher payload mass appears to coincide with a higher success rate
- The majority of launches with payload mass over 8000 kg were successful
- KSC LC 39A had a 100% success rate with payload mass under ~5500 kg

INSIGHTS: EDA - DATA VISUALIZATION

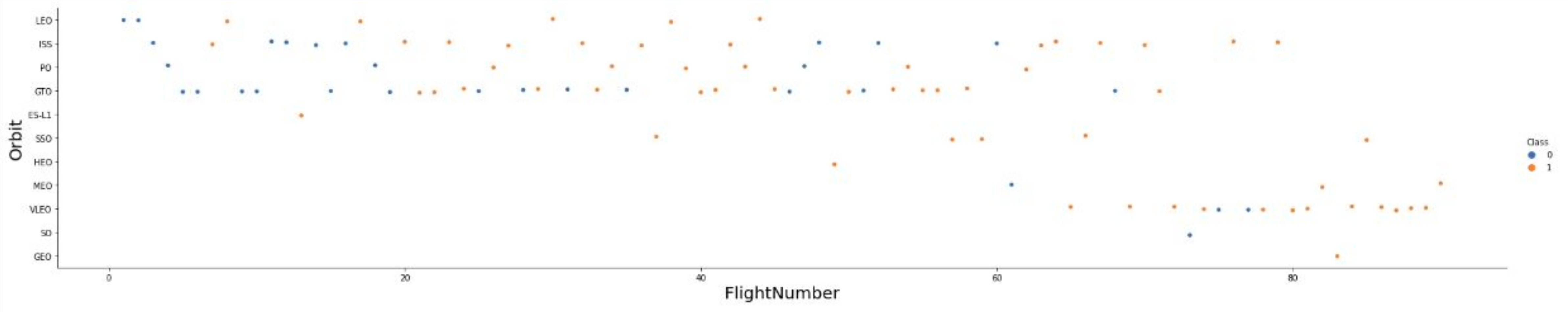
Orbit Type x Success Rate



- ES-L1, GEO, HEO, and SSO all had a 100% success rate
- SO had a 0% success rate

INSIGHTS: EDA - DATA VISUALIZATION

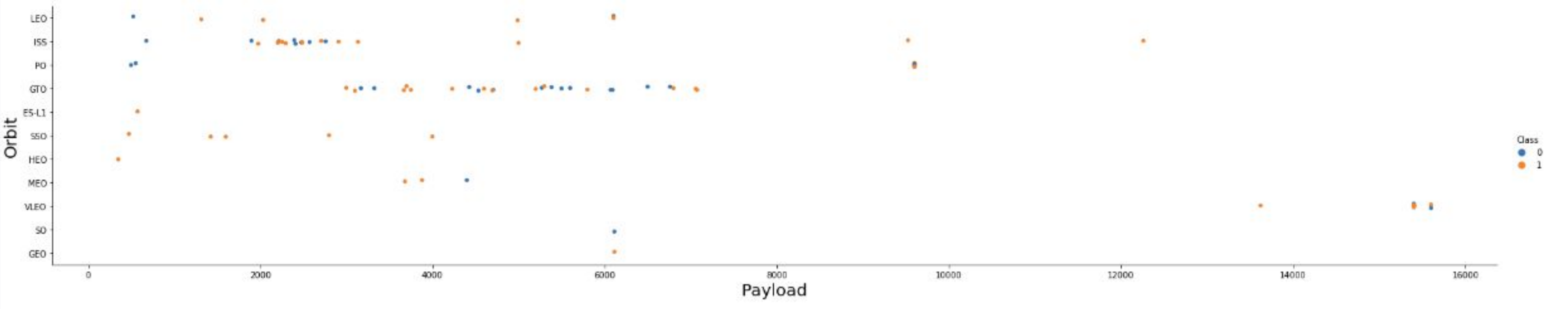
Flight Number x Orbit Type



- LEO showed a general increase in success rate as the number of flights increased
- VLEO showed a high rate of success at a larger number of flights

INSIGHTS: EDA - DATA VISUALIZATION

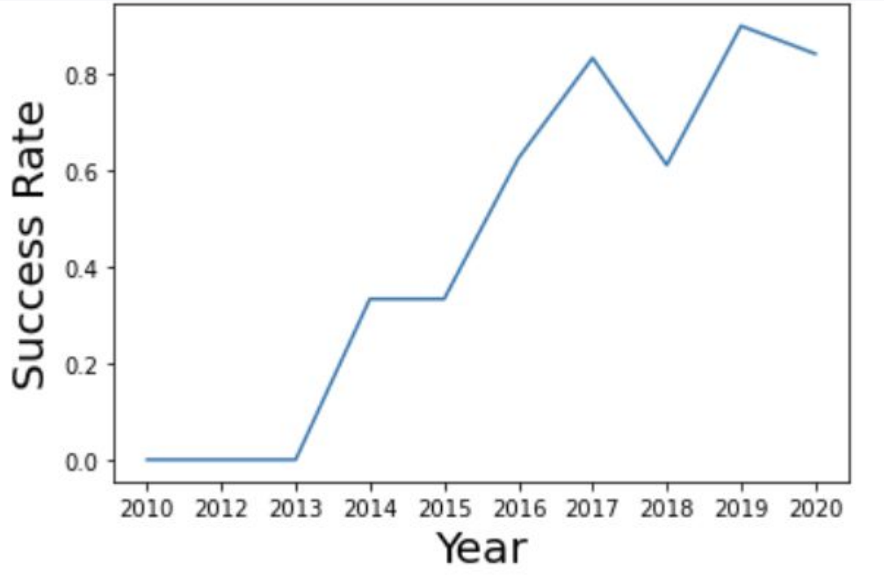
Payload x Orbit Type



- GTO appeared to have a lower rate of success at higher payloads (>5000 kg)
- SSO had a 100% success rate at payloads less than 4000 kg

INSIGHTS: EDA - DATA VISUALIZATION

Launch Success Yearly Trend



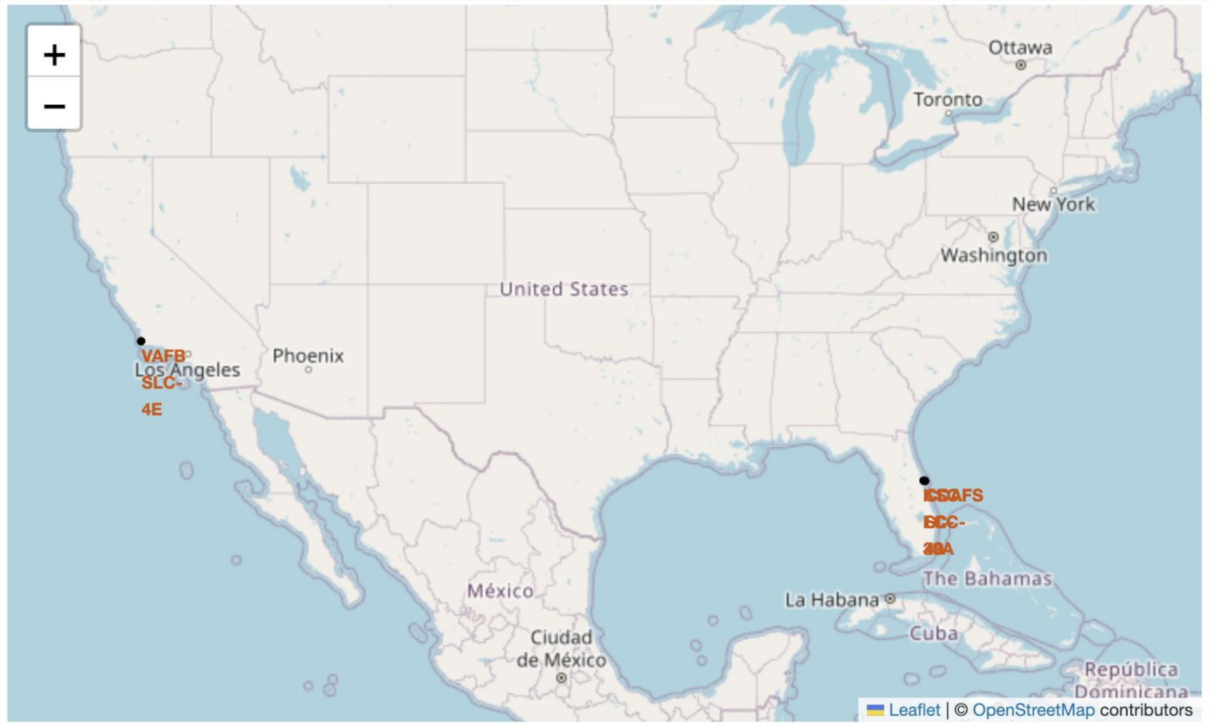
- Overall, success rate appeared to increase year over year
- 2014 to 2015 showed little to no increase in success rate
- Success rate decreased from 2017 to 2018 and from 2019 onwards

Section 3

Launch Sites Proximities Analysis

LAUNCH SITES LOCATIONS ANALYSIS

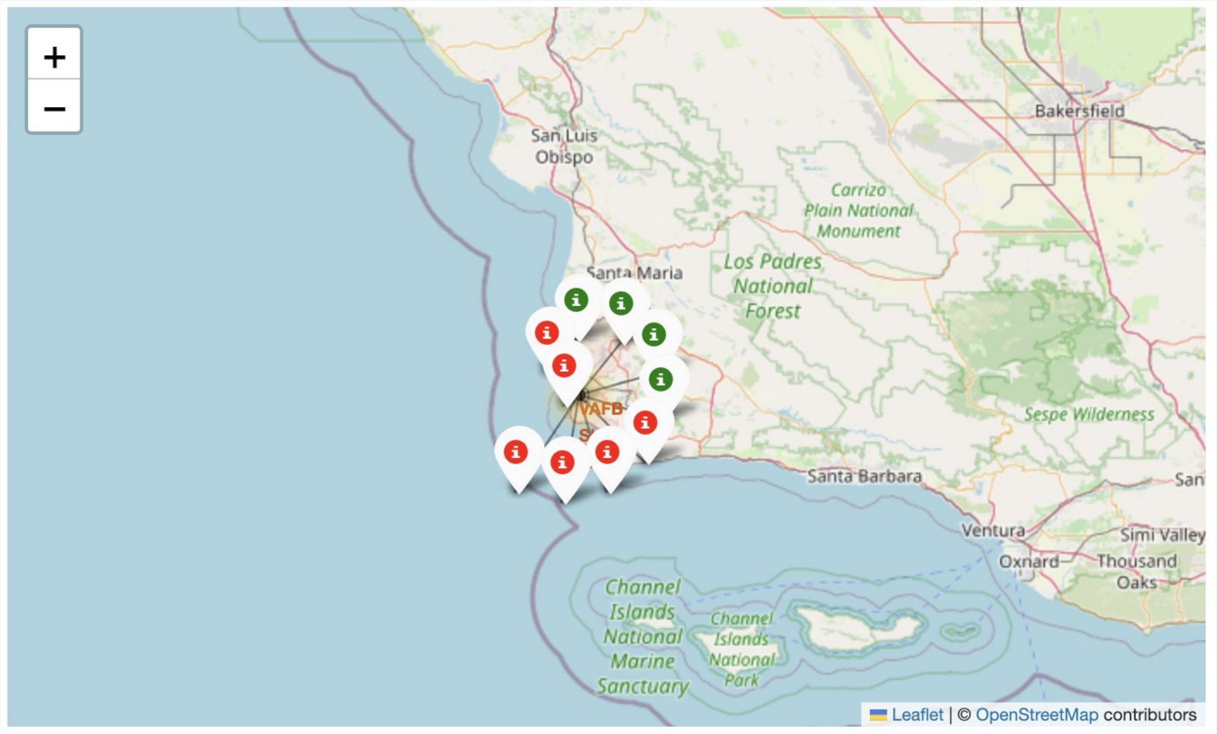
Launch Site Location Markers



The map above shows the locations (marked in red) of the launch sites. Note that these locations are as far south as possible (in the U.S.) and are close to the coast.

LAUNCH SITES LOCATIONS ANALYSIS

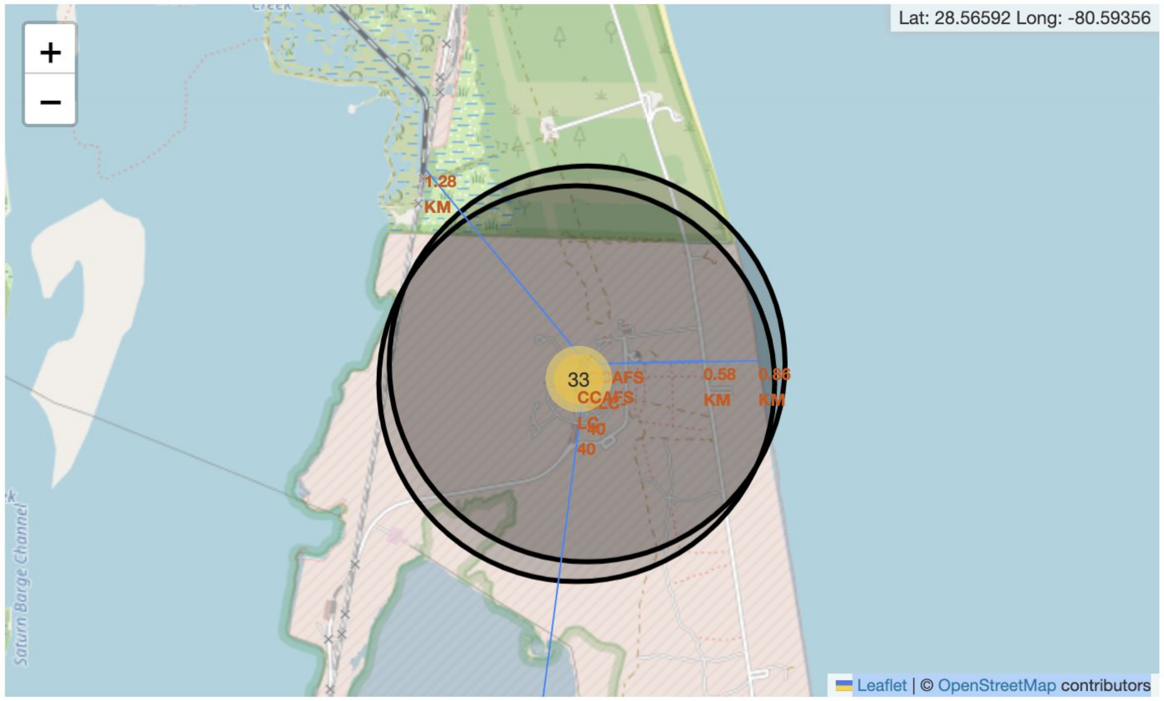
Color-Labeled Launch Outcomes



The map above shows the outcomes of launches at VAFB SLC 4E. Red represents an unsuccessful landing, while green represents a successful landing.

LAUNCH SITES LOCATIONS ANALYSIS

Launch Site Proximities



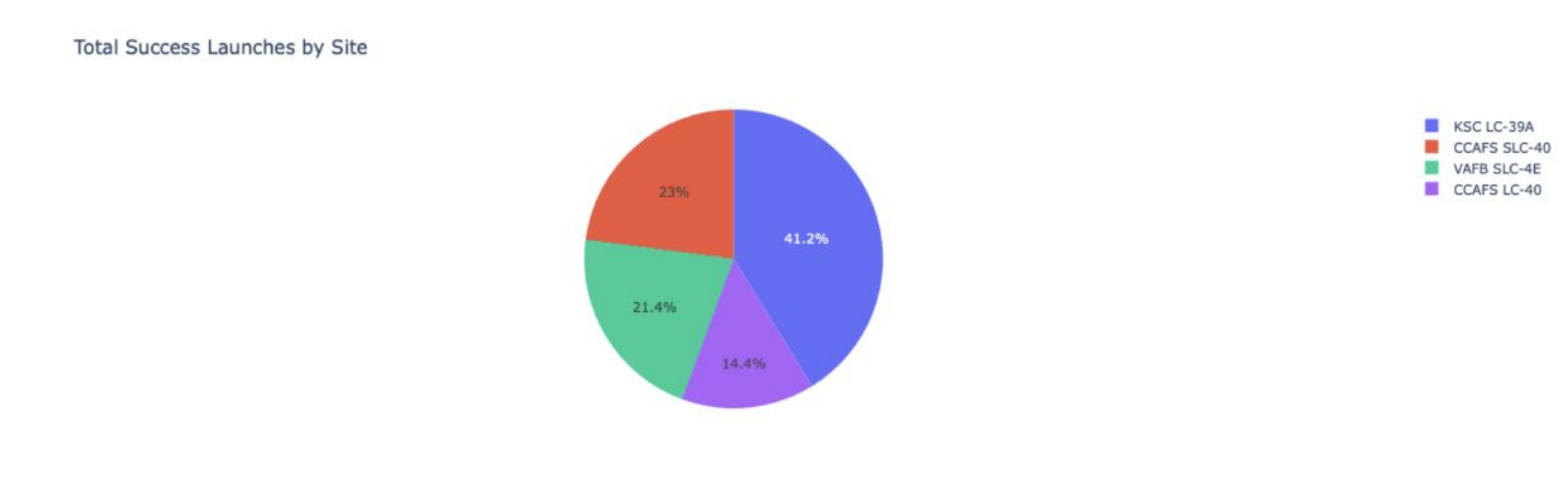
The map above shows the proximity of the CCAFS SLC 40 launch site to several landmarks, such as the closest railway (1.28 km), highway (0.58 km), and coastline (0.86 km).

Section 4

Build a Dashboard with Plotly Dash

LAUNCH RECORDS DASHBOARD

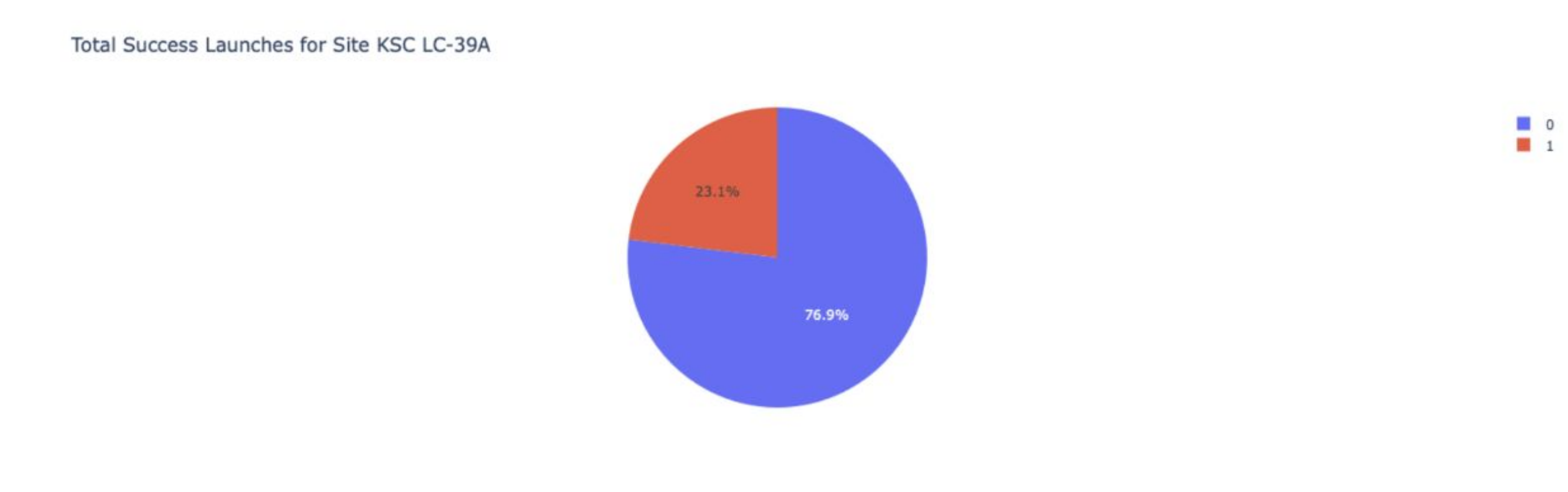
Launch Success Count for all Sites



- The pie chart above shows the launch success count for all launch sites
- Here, we can see that KSC LC-39A had the most successful launches

LAUNCH RECORDS DASHBOARD

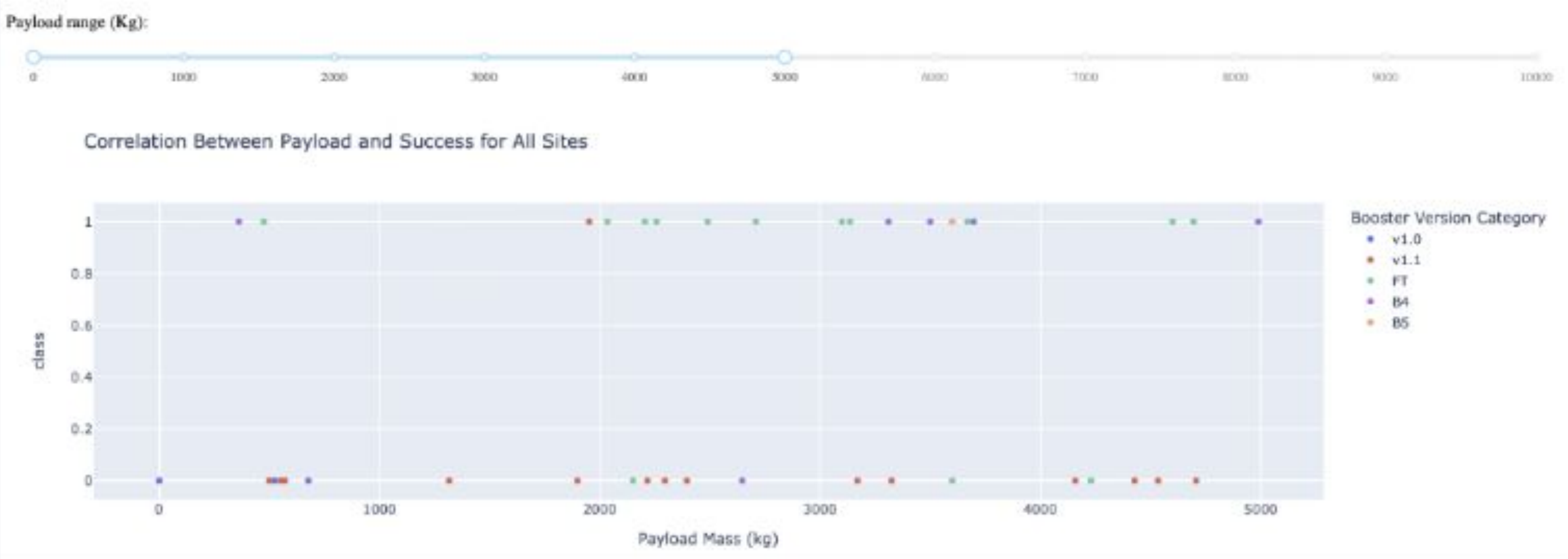
Launch Site with the Highest Launch Success Ratio



- The pie chart above shows the launch success ratio for KSC LC-39A
- Here, we can see that this site had a 76.9% success rate

LAUNCH RECORDS DASHBOARD

Payload x Launch Outcome for all Launch Sites

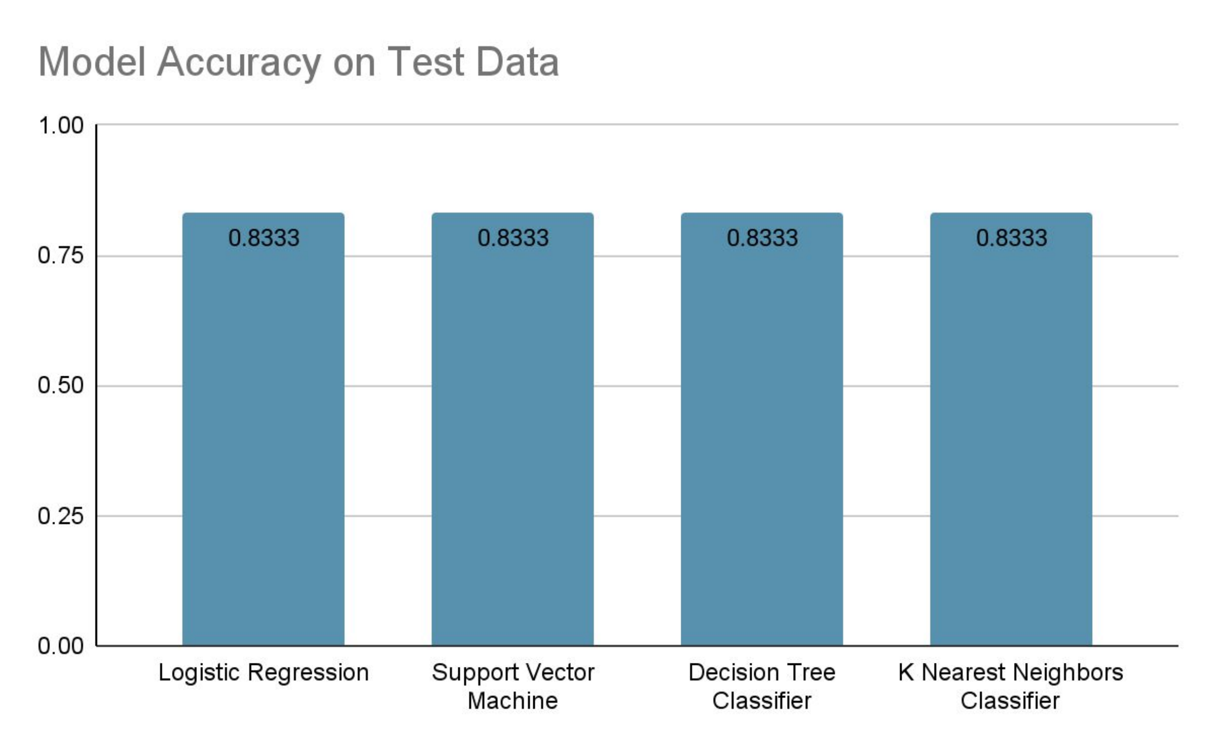


- The chart above shows the relationship between Payload Mass and Outcome
- Payloads less than 5300 kg tended to have a higher success rate
- Payloads greater than 5300 kg tended to have lower success rate

Section 5

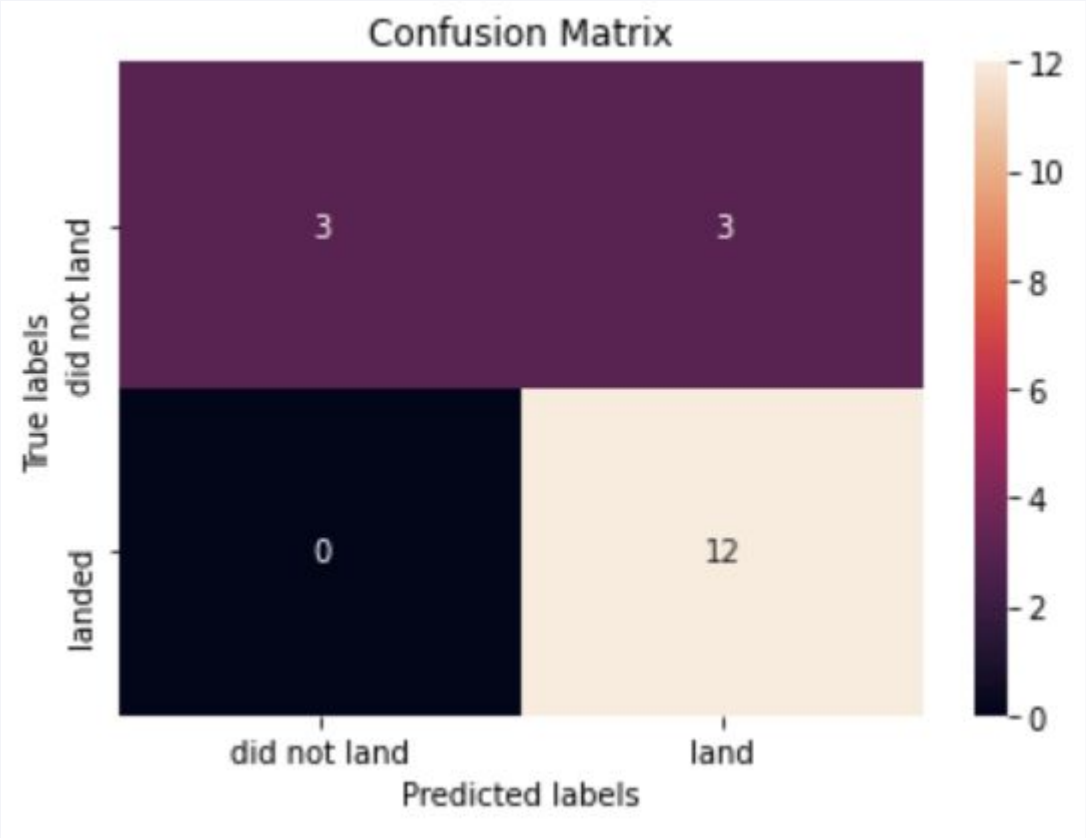
Predictive Analysis (Classification)

CLASSIFICATION ACCURACY



The chart above shows that each model yielded an accuracy score of ~83.33%

CONFUSION MATRIX



The confusion matrix above represents the performance of each model, as they all performed identically. Here, we can see that each model correctly yielded a true positive 12 times, and a true negative 3 times. However, the model also yielded 3 false positives.

CONCLUSION

- All models yielded the same accuracy on the test set (~83.33%)
  - The Decision Tree Classifier yielded the highest accuracy on the training set (~88.88%)
- SpaceY can predict a successful landing of a SpaceX first stage booster with 83.33% accuracy
- Potential future work can be done to improve this accuracy
  - Further collection of data to increase training and testing dataset sizes
  - Revisit the parameter tuning of the Decision Tree Classifier, as it yielded the highest accuracy on the training set
  - Employ more powerful classifier models, such as LightGBM

APPENDIX

- Notebooks used to complete this project:
  - https://github.com/ConradKleykamp/IBM-Applied-Data-Science-Capstone/blob/main/1-Data-Collection-API.ipynb
  - https://github.com/ConradKleykamp/IBM-Applied-Data-Science-Capstone/blob/main/2-Data-Collection-with-Web-Scraping.ipynb
  - https://github.com/ConradKleykamp/IBM-Applied-Data-Science-Capstone/blob/main/3-Data-Wrangling.ipynb
  - https://github.com/ConradKleykamp/IBM-Applied-Data-Science-Capstone/blob/main/4-EDA-with-SQL.ipynb
  - https://github.com/ConradKleykamp/IBM-Applied-Data-Science-Capstone/blob/main/5-EDA-with-Visualization.ipynb
  - https://github.com/ConradKleykamp/IBM-Applied-Data-Science-Capstone/blob/main/6-Launch-Sites-Locations-Analysis-with-Folium.ipynb
  - https://github.com/ConradKleykamp/IBM-Applied-Data-Science-Capstone/blob/main/7-Interactive-Dashboard-with-Plotly-Dash.py
  - https://github.com/ConradKleykamp/IBM-Applied-Data-Science-Capstone/blob/main/8-Machine-Learning-Prediction.ipynb

Thank you!