IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Tobi Ibidapo-Obe

December, 2023

Outline

Executive Summary

Objective

We (SpaceY) want to use SpaceX's historical data to employ cutting-edge predictive models and unravel the complexities behind rocket launch outcomes to conduct safer and more cost-efficient space missions.

Methodologies

- Data Collection via API and Web scraping
- Data Wrangling
- EDA with SQL and Python libraries (Pandas, Matplotlib)
- Interactive Visual Analytics and Dashboard with Folium and Plotly Dash
- Predictive Analysis using Classification modeling and Grid Search for algorithm optimization

Results and Next Steps

- We created a predictive model with high accuracy to predict the Falcon9 rocket launch outcome of the first stage. Our predictive model and data analytics revealed patterns that can reliably estimate the cost of a rocket launch. We can leverage the potential of data science in the space launch domain to optimize rocket launch success rates and propel advancements within the aerospace industry. Continued research, collaboration, and refinement of predictive models are crucial for further enhancing the reliability and precision of rocket launch outcome predictions in making price estimates for rocket launch.

Introduction

Project background and context

- We (SpaceY) want to launch rockets with minimal costs.
- SpaceX advertises Falcon 9 rocket launches on its website at 62 million dollars, while other providers cost upwards of 165 million dollars. Much of the savings is because SpaceX can reuse the first stage of its rocket launch. Therefore, if we (SpaceY) can predict the landing outcome of the first stage, we can determine the cost of a rocket launch and make SpaceY pricing very competitive.

Problems we want to find answers

- Determine the price of each rocket launch
- Determine if SpaceX will reuse the first stage

Source: miro.medium.com

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - We used SpaceX API and Web scraping of a Wikipedia page to get SpaceX historical data
- Performed data wrangling
  - We performed some EDA to find some patterns in the data
  - We converted the outcome categorical variable into two-class label (0: Unsuccessful; 1: Successful)
- Performed exploratory data analysis (EDA) using visualization and SQL
- Performed interactive visual analytics using Folium and Plotly Dash
- Performed predictive analysis using classification models
  - We built, evaluated, improved, and found the best classification model using the accuracy score

Data Collection

Data were collected using SpaceX API and a Wikipedia page

- The API end points shows data about SpaceX rocket launches, including information about the rocket used, payload delivered, launch specifications (such as dates of launch, launch site), landing specifications, and landing outcome.

- Web scraped some HTML tables that contained valuable Falcon 9 launch records from a Wikipedia page.

Data Collection – SpaceX API

Data collection with SpaceX REST calls was performed in 3 major steps/tasks.

Completed SpaceX API calls notebook: https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone/blob/48ec78aaf728bb09ecbbb6a2ef20eb5dbe2b7970/1_Data-Collection-and-Wrangling/SpaceX_API-Data.ipynb

3 Tasks Performed

We made a GET request to parse the SpaceX rocket launch data from SpaceX API

- Used a defined series of **helper functions** to extract information in the rocket launch data; getBoosterVersion, getPayLoadData, getLaunchSite, getCoreData
- From the base API, I created a JSON object to make the API results more consistent and converted the JSON result to a dataframe (data), then filtered to specific features (date, rocket, payloads, launchpad, cores, flight_number)
- From four columns in the dataframe (data), I applied the Helper functions to retrieve vital information; 'rocket' column - getBoosterVersion, 'payload' column - getPayLoadData, 'launchpad' column – getLaunchSite, 'cores' column – getCoreData
- I created a list of variables to construct a new list dataframe (launch_dict) which included the created list of variables and two columns (flight_number and date) from the dataframe (data). I transformed the new dataframe into a pandas dataframe (df)

We filtered the dataframe (df) to keep only the Falcon 9 rocket launches into a new dataframe (data_falcon9) and checked for missing values.

We imputed the mean into the missing values of the 'PayloadMass' column and saved the dataframe(data_falcon9) as dataset_part_1.csv

Data Collection - Scraping

Data collection from Wikipedia page was performed in 3 major steps/tasks.

Completed SpaceX web scraping notebook:
https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone/blob/48ec78aaf728bb09ecbbb6a2ef20eb5dbe2b7970/1_Data-Collection-and-Wrangling/SpaceX_Web-scraping.ipynb

3 Tasks Performed

We performed an HTTP GET method to request the Falcon9 rocket launch HTML page from the Wikipedia page:
https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

- We defined a series of helper functions (date_time, booster_version, landing_status, get_mass, extract_column_from_header) to process web scraped HTML table.

↓

We extracted all variable names from the HTML table header.

↓

We created a dataframe (launch_dict) by parsing the launch HTML tables and saved as spacex_web_scraped.csv

Data Wrangling

We performed some Exploratory Data Analysis (EDA) to find some patterns in the data and then created the numeric label for launch outcomes to train classification models. Patterns from the EDA include;

Number of launches on each site

```
df['LaunchSite'].value_counts()

Out[5]: LaunchSite
        CCAFS SLC 40    55
        KSC LC 39A      22
        VAFB SLC 4E     13
        Name: count, dtype: int64
```

Number and occurrence of each orbit

```
df['Orbit'].value_counts()

Out[6]: Orbit
        GTO      27
        ISS      21
        VLEO     14
        PO        9
        LEO       7
        SSO       5
        MEO       3
        ES-L1     1
        HEO       1
        SO        1
        GEO       1
        Name: count, dtype: int64
```

Number and occurrence of mission outcome per orbit type

```
landing_outcomes= df['Outcome'].value_counts()
landing_outcomes

Out[7]: Outcome
        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: count, dtype: int64
```

Completed data wrangling notebook: https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone/blob/48ec78aaf728bb09ecbbb6a2ef20eb5dbe2b7970/1_Data-Collection-and-Wrangling/SpaceX_Data-Wrangling.ipynb

EDA with Data Visualization

- **Visualized the relationship between Flight Number and Launch Site**
  Plotted a scatter plot chart to understand how launch outcome affect the distribution of flight numbers across different launch sites.
- **Visualized the relationship between Payload and Launch Site**
  Used a scatter plot chart to examine how launch outcomes affect payload masses at different launch sites.
- **Visualized the relationship between success rate of each orbit type**
  Explored with a bar chart the success rates of different orbit types to identify which orbits tend to have higher success rates.
- **Visualized the relationship between Flight Number and Orbit type**
  Plotted a scatter plot chart to investigate how flight numbers and orbit types relate to launch outcomes.
- **Visualized the relationship between Payload and Orbit type**
  Explored with a scatter plot chart the relationship of launch outcomes with payload masses and orbit types.
- **Visualized the launch success yearly trend**
  Plotted a line chart to identify the trend in the success rate of launches over the years.

Completed data visualization notebook: https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone/blob/48ec78aaf728bb09ecbbb6a2ef20eb5dbe2b7970/2_Exploratory-Data-Analysis/EDA-dataviz.ipynb

EDA with SQL

SQL tasks performed

- Retrieved the names of unique launch sites.
- Displayed 5 records where launch sites begin with the string ‘CCA’.
- Displayed the total payload mass carried by boosters launched by NASA (CRS).
- Displayed average payload mass carried by booster version F9 v1.1
- Retrieved the date when the first successful landing outcome in ground pad was achieved.
- Retrieved the total number of successful and failure mission outcomes.
- Retrieved the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
- Displayed names of the booster_versions which have carried the maximum payload mass.
- Retrieved records for the month names, failure landing_outcomes in drone ship ,booster versions, launch_site in year 2015.
- Ranked the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Completed SQL notebook: https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone/blob/48ec78aaf728bb09ecbbb6a2ef20eb5dbe2b7970/2_Exploratory-Data-Analysis/EDA-sql.ipynb

Interactive Map with Folium

Created and added map objects to a folium map

- Marked all launch sites on a map
  - Represented each launch site with a circle object. These markers (circles) help to locate and visualize the launch sites for easy identification.
- Marked the success/failed launches for each site
  - Added MarkerCluster to identify similar launch sites and differentiated between successful and failed launches using different colors to depict success and failure (Green for success and Red for failure). This was useful in visualizing the outcome at each site to identify patterns and trends in launch outcomes.
- Calculated the distances between a launch site to its proximities
  - Connected launch sites with proximities (relevant points like coastlines) using PolyLines to draw lines between launch sites and proximity points to visualize the distances. This offered insights into accessibility or geographical constraints.

Completed Folium map notebook: https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone/blob/48ec78aaf728bb09ecbbb6a2ef20eb5dbe2b7970/3_Data-Visualization/DataViz-Folium.ipynb

Visual Analytics Dashboard with Plotly Dash

Built an interactive real-time visual analytics dashboard

- **Visualized the successful launch outcome at each site**
  - Plotted a pie chart to understand the distribution of launch outcomes at different launch sites, indicating the site with the largest successful launch and highest success rate.
- **Visualized the relationship between Payload and Launch Success Outcome**

  Used a scatter plot chart to examine how payload masses vary with successful launch outcomes to determine the likelihood of a successful launch with the weight of the rocket payloads.

Completed Plotly map notebook: https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone/blob/48ec78aaf728bb09ecbbb6a2ef20eb5dbe2b7970/3_Data-Visualization/DataViz-Plotly.py

Predictive Analysis (Classification)

Classification Model

Model Development:

- We created 2 classes
- Standardized the data
- Split data into training and test data
- Created four classification algorithms (Logistic regression, SVM, Decision Trees and K-nearest neighbors)

Model Evaluation:

- We used GridSearchCVto find the best hyperparameter.
- We calculated the accuracy score on the test data for the four algorithms.
- Used confusion matrix to comprehensively asses the models performance.

Selected the Best Model

- We used the highest accuracy score and confusion matrix to pick the best model.

Completed Predictive Analysis notebook: https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone/blob/48ec78aaf728bb09ecbbb6a2ef20eb5dbe2b7970/4_Predictive-Analysis/SpaceX-Machine_Learning_Prediction.ipynb

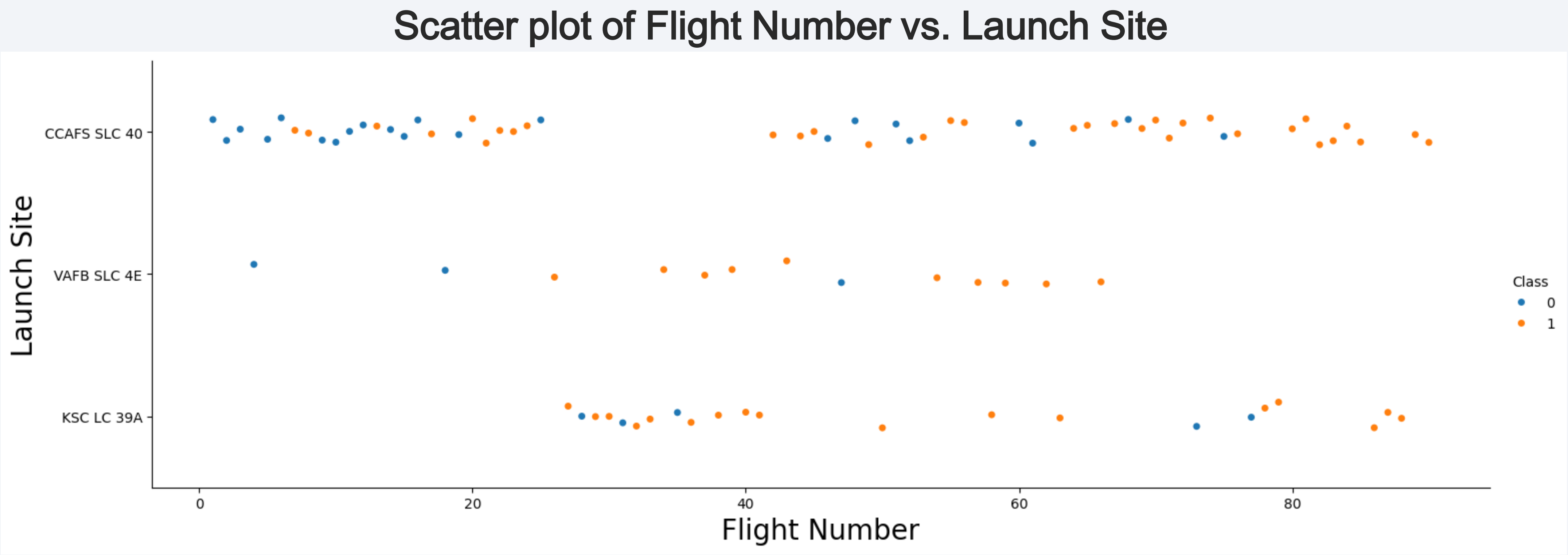
Results

- Exploratory Data Analysis
- Screenshots of Interactive analytics and Dashboards
- Predictive analysis

Section 2

Insights drawn from EDA

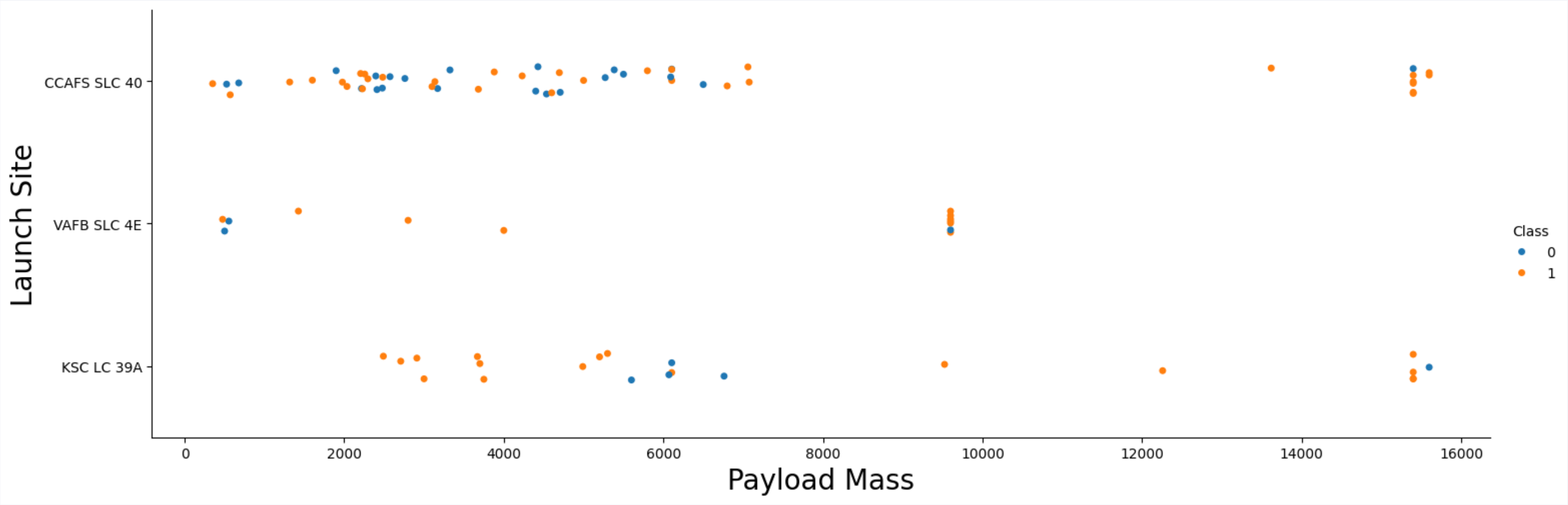
Flight Number vs. Launch Site



We see that as the flight number increases, the first stage is more likely to land successfully. Different launch sites have different success likelihood. KSC LC-39A seem to have higher likelihood the first stage will return or be successful.

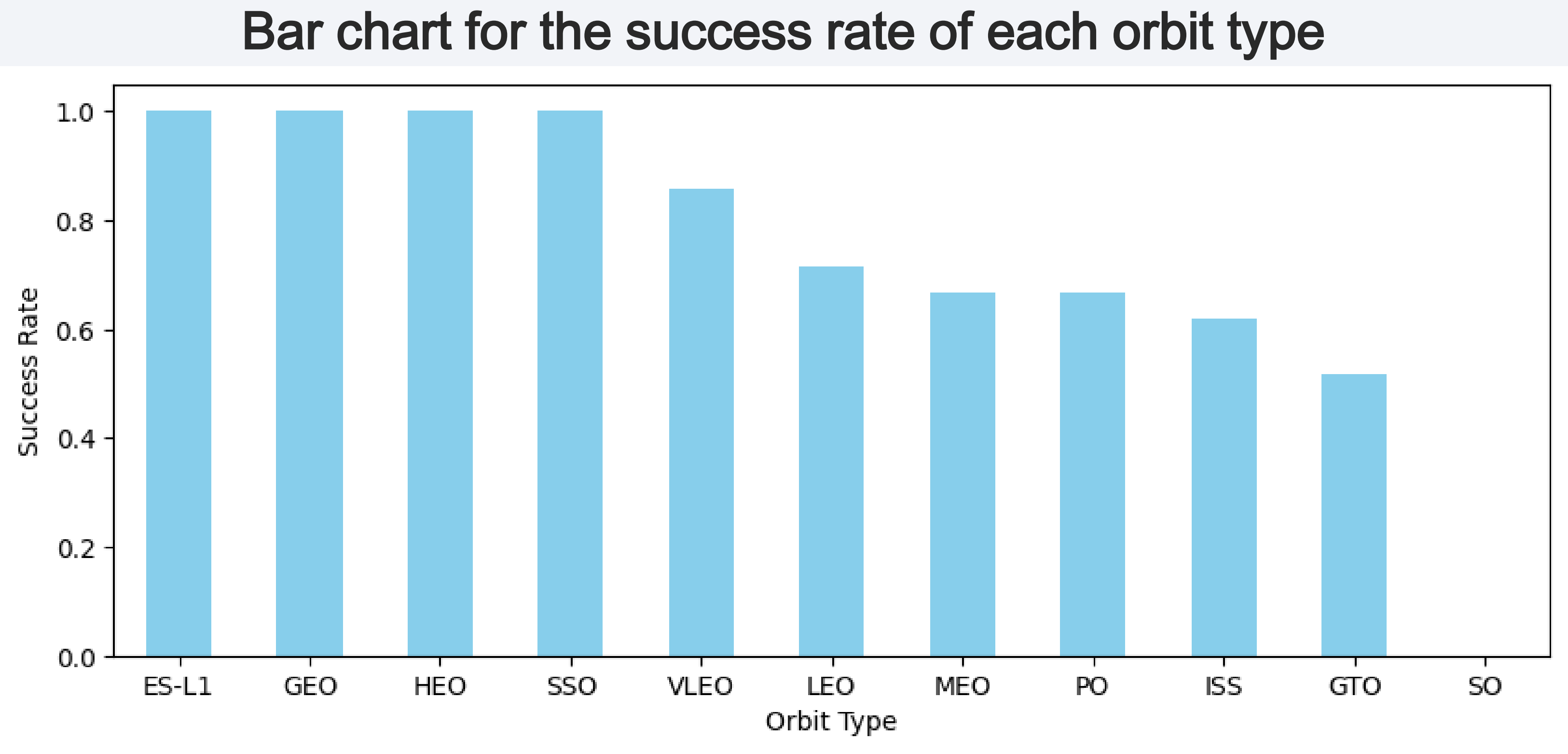
Payload vs. Launch Site

Scatter plot of Payload vs. Launch Site



It seems the more massive the payload, the less likely the first stage will return. KSC LC-39A seem to have lower payload mass deployed at the site and hence higher likelihood the first stage will return or be successful, while no heavy payload mass (greater than 10,000kg) deployed for the VAFB-SLC launch site.

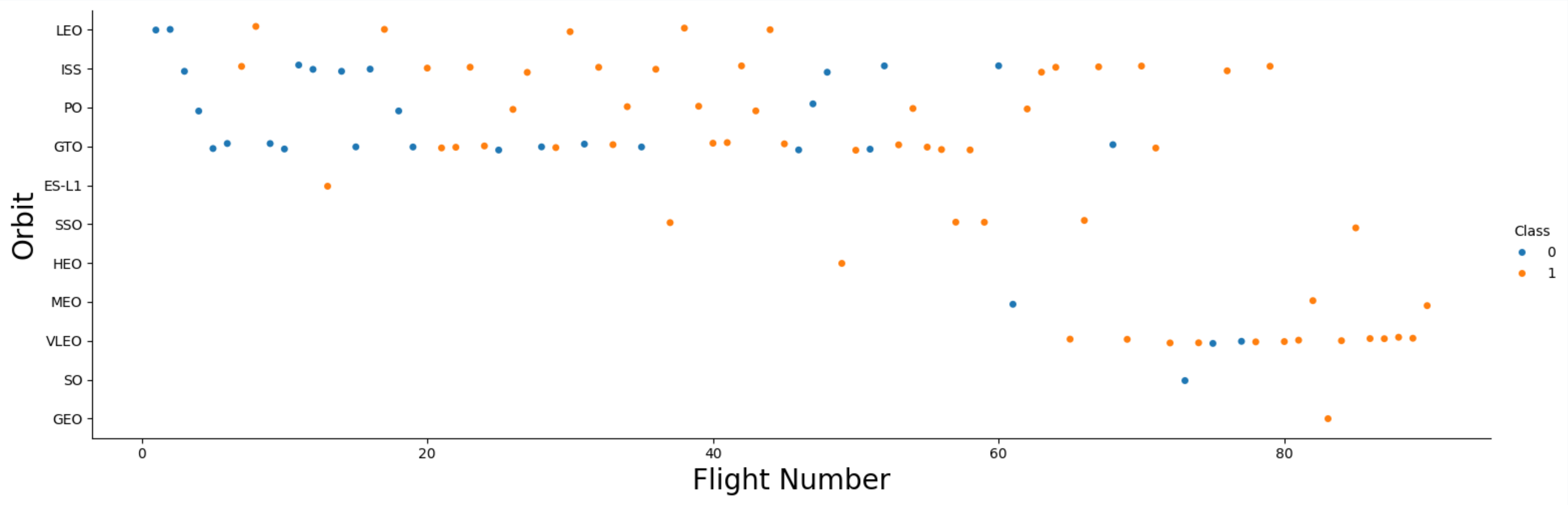
Success Rate vs. Orbit Type



Four orbits (ES-L1, GEO, HEO and SSO) have absolute success (100%) for the first stage to land successfully, and a stark contrast of absolutely no likelihood (0%) of success rate at orbit SO.

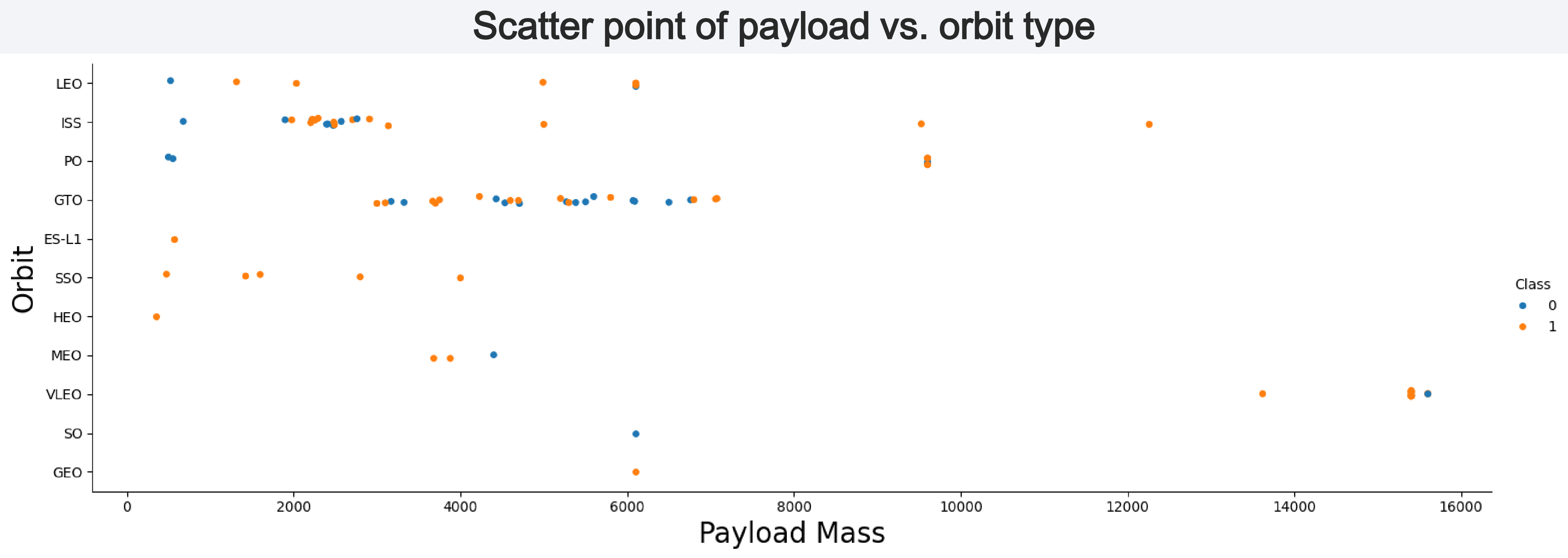
Flight Number vs. Orbit Type

Scatter point of Flight number vs. Orbit type



As the flight number increases, the first stage is more likely to land successfully. The orbits where more launch attempts (flight number) are taken also show high likelihood for successful first stage landing as seen in the LEO orbit. However, there seems to be no relationship between flight number when in GTO orbit.

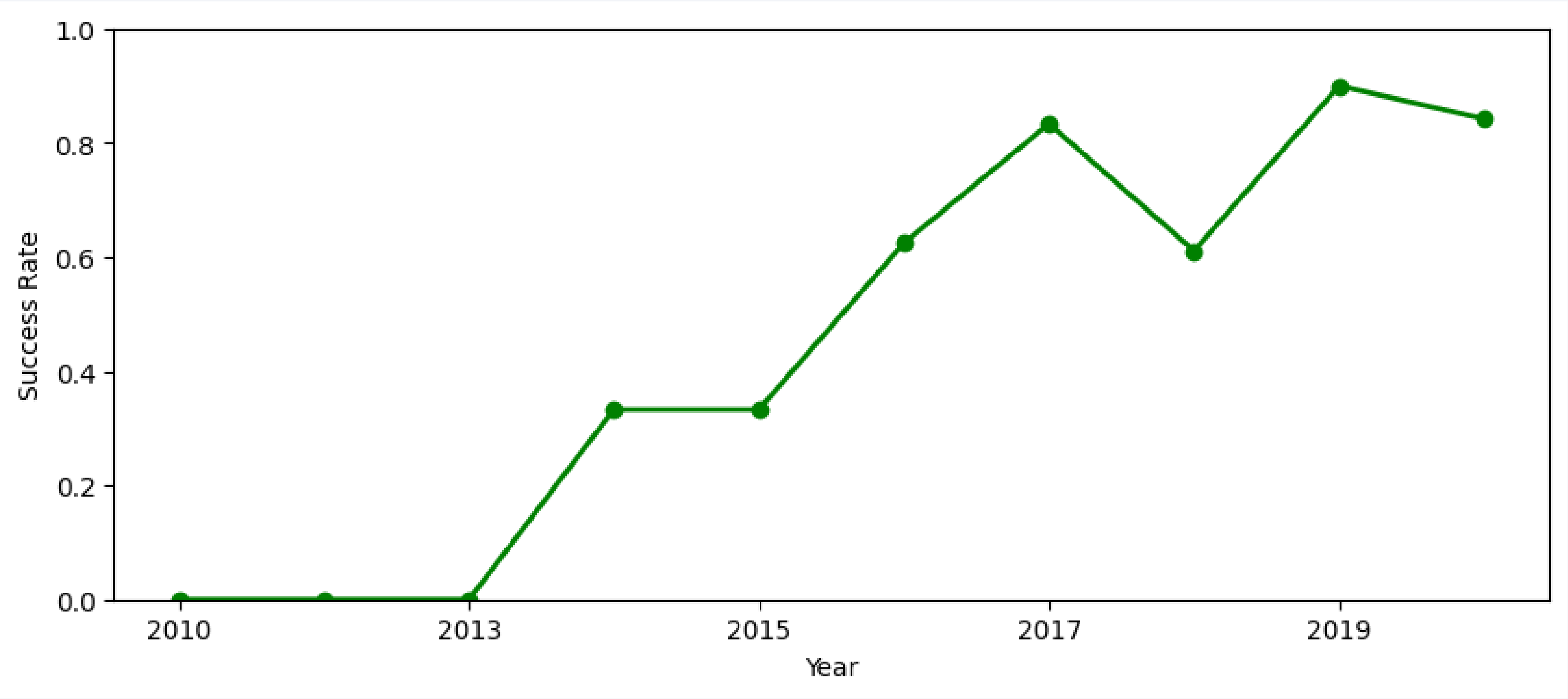
Payload vs. Orbit Type



While more massive the payload shows a less likely first stage return, lesser payload mass deployed at an orbit sets up successful rocket return (e.g Orbits ES-L1 and HEO have payloads less than 1,000kg and the maximum payload at orbit SSO is 4,000kg all having successful first stage return. Heavy payloads typically show successful landing rate at Polar orbits (LEO and ISS). However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are seen.

Launch Success Yearly Trend

Line chart of yearly average success rate



Success rate has been increasing since 2013.

All Launch Site Names

Names of the unique launch sites

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

Only four sites or locations were used to carry out the Falcon 9 rocket launches

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA`

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Though two launch sites begin with the string 'CCA' (CCAFS LC-40 and CCAFS SLC-40). The result query for the first 5 records returned shows only CCAFS LC-40, indicating more activities at this launch site.

Total Payload Mass

Total payload carried by boosters from NASA

Total_payload_mass
45596

Total payload carried by boosters from NASA = 45596 kg signifies that NASA actively uses SpaceX rockets for its space missions and seem to be SpaceX biggest customer.

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1

Average_Payload_Mass
2928.4

Average payload carried by booster version F9 v1.1 = 2928.4 kg signifies that the average cargo, satellites, or scientific instruments that this booster could transport into space per mission was small compared to its latest booster versions.

First Successful Ground Landing Date

Dates of the first successful landing outcome on ground pad

First_successful_landing
2015-12-22

SpaceX achieved its first successful ground landing of the Falcon 9 rocket on December 22, 2015, marking a significant milestone in SpaceX's efforts to develop reusable rocket technology.

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

These boosters have been able to carry payload mass of up to 6,000 kg and successfully landed on a drone ship in space.

Total Number of Successful and Failure Mission Outcomes

Total number of successful and failure mission outcomes

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

The total number of successful missions significantly outweighs the number of failures. Majority of the space missions were successful (100) out of 101.

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass

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

These boosters (majorly the B5 boosters) carried the heaviest payloads

2015 Launch Records

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Month	Failure_Landing_Outcome	Booster_Version	Launch_Site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

On January 10, 2015 Falcon 9 v1.1 attempted a drone ship landing after delivering cargo to the International Space Station (ISS). The landing was close but unsuccessful due to a loss of hydraulic fluid just before landing, causing a hard landing on the drone ship. Similar occurrence happened in April of the same year, the rocket reached the drone ship but landed too hard, resulting in a failed landing.

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

Count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

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

The table above shows the landing outcomes between June 2010 and March 2017.

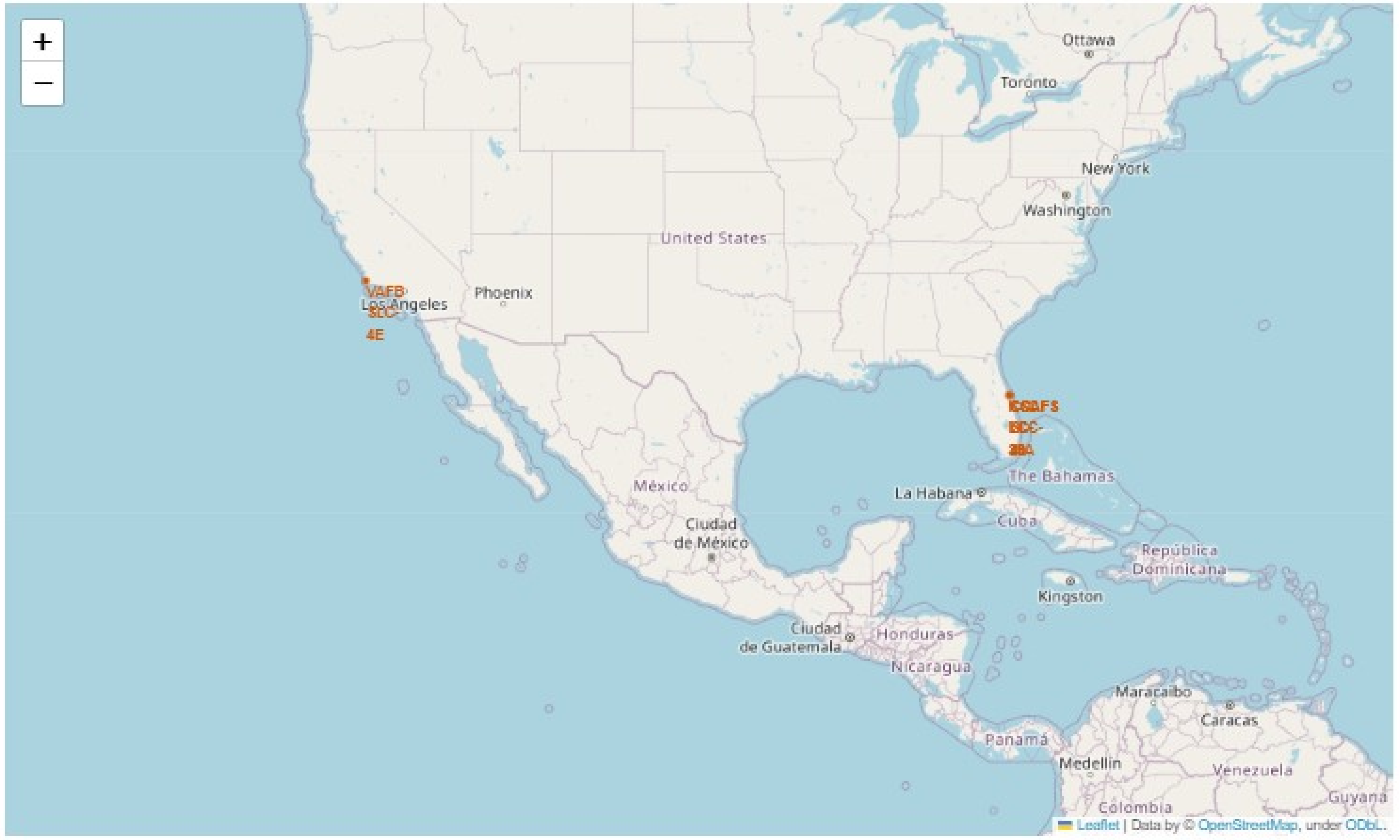
Section 3

Launch Sites Proximities Analysis

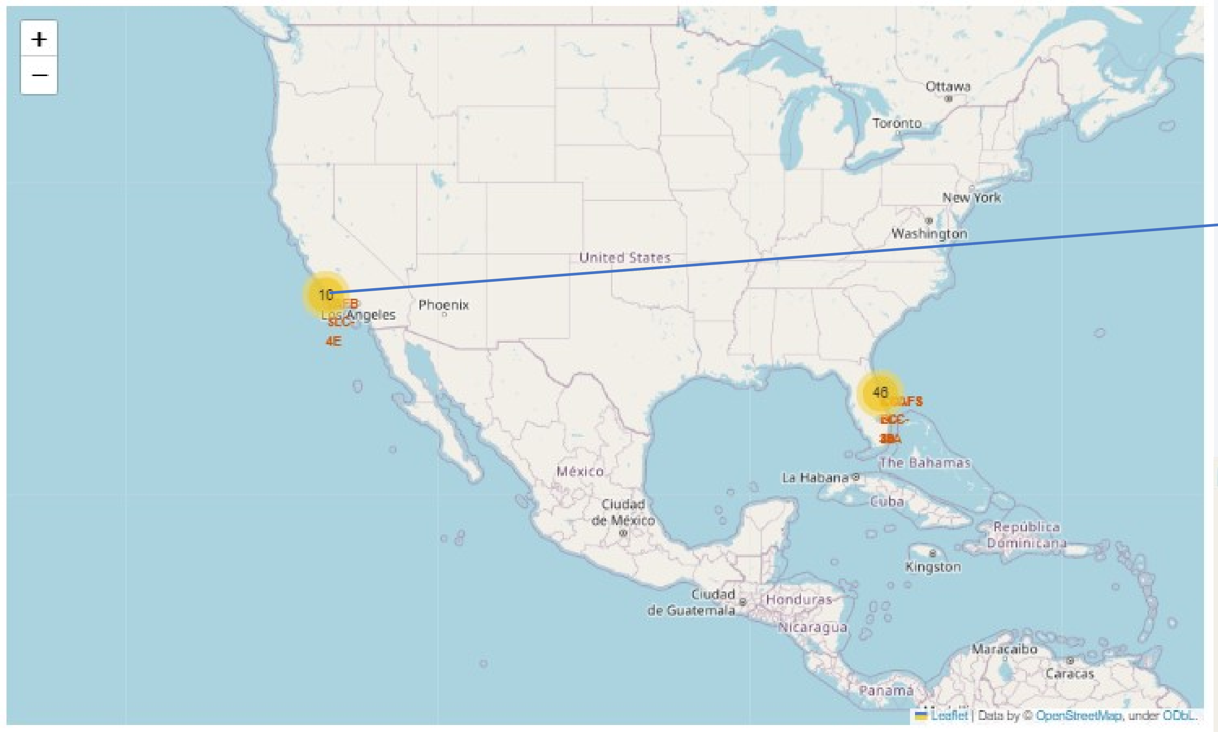
Map of all Launch sites

Three of the launch sites are situated at latitude of approx. 28.5 degrees North, which is not close to the Equator. The fourth site (VAFB SLC-4E) is at a higher latitude of around 34.6 degrees North, farther from the Equator at the west, precisely California compared to the Florida launch sites.

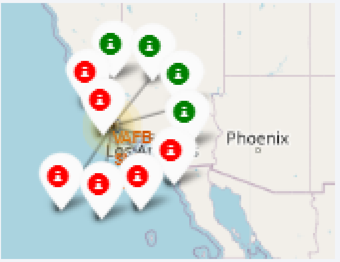
The launch sites were located near coastlines due to various logistical, safety, and operational advantages associated with coastal launch sites.



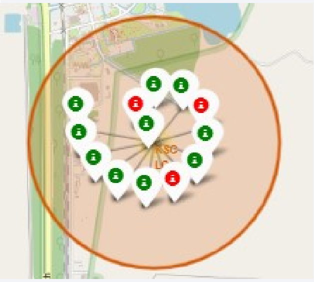
Launch Outcomes for each Site



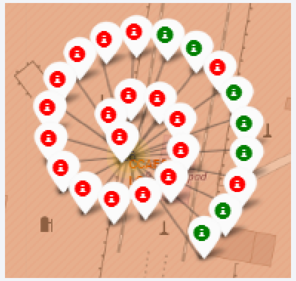
VAFB SLC-4E



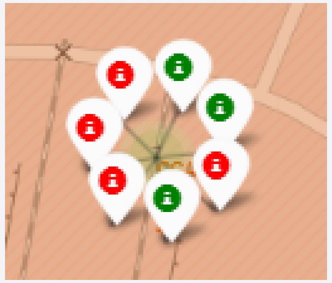
KSC LC-39A



CCAFS LC-40

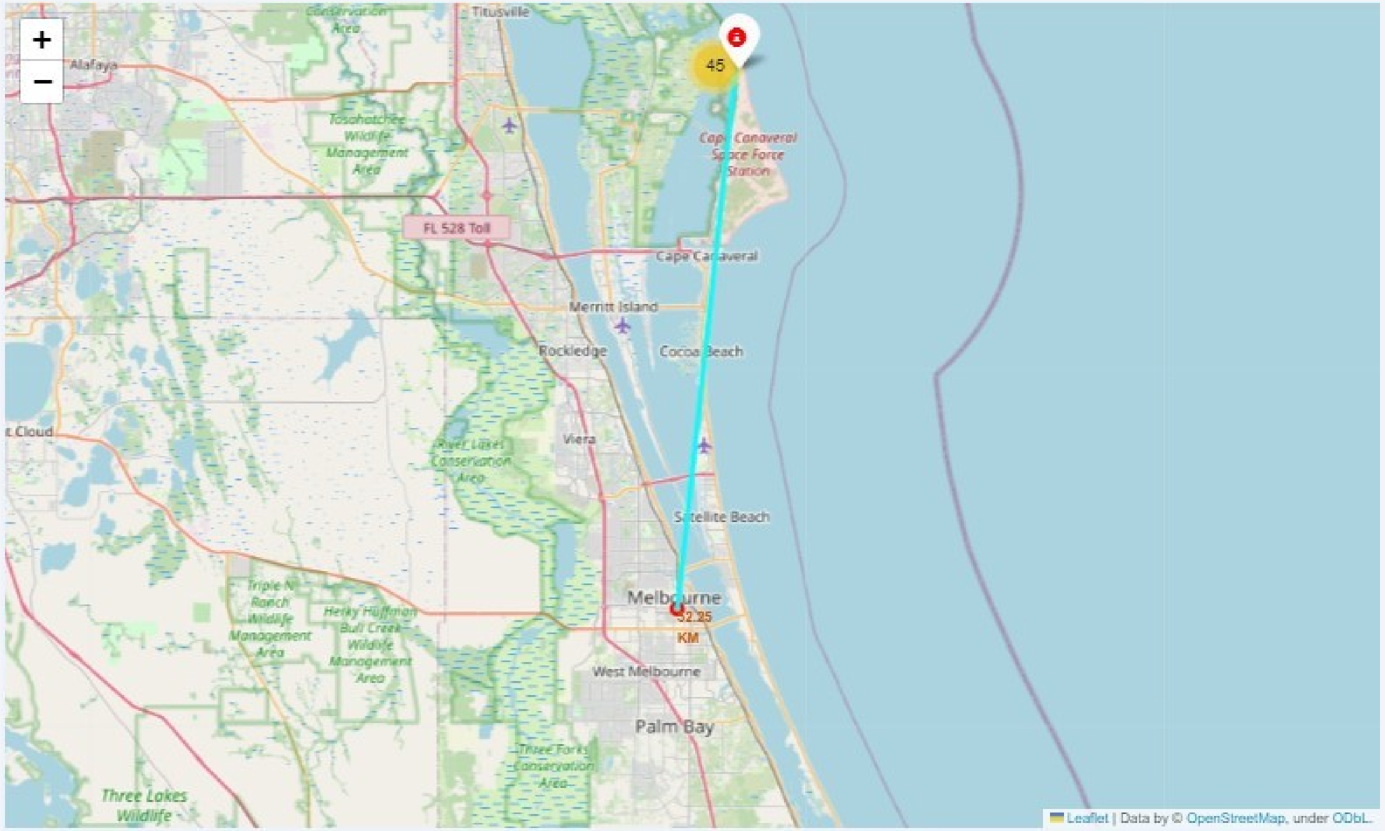


CCAFS SLC-40



At VAFB SLC-4E, 4 successful out of 10. At KSC LC-39A, 10 out of 13. At CCAFS LC-40, 7 out of 26. At CCAFS SLC-40, 3 out of 7

Distance of Launch Site to nearby City



Cape Canaveral launch site (CCAFS SLC-40) is relatively nearer to Melbourne compared to other sites proximity to the nearest city. The CCAFS SLC-40 site might be relatively closer compared to other major launch sites, the distance between Melbourne and the launch site is still considerable and involves significant travel distance by air or Sea, about 52 kilometers.

Section 4

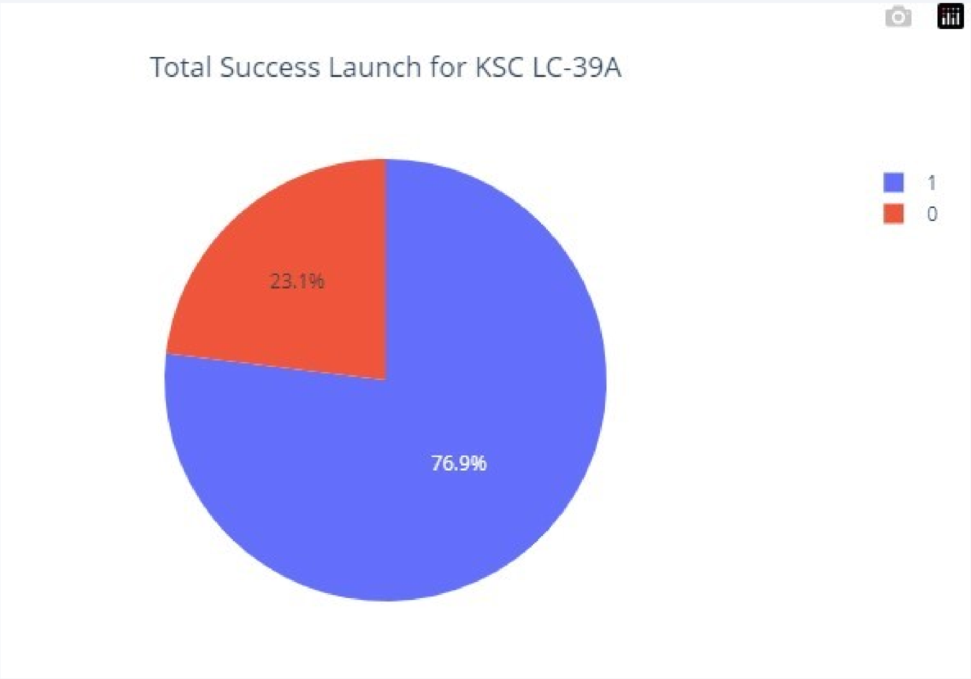
Build a Dashboard with Plotly Dash

Successful Launch Outcome for All Sites

KSC LC-39A had the largest successful launches, 10 successful landing outcomes out of 13.

Launch Site with the highest Successful outcome

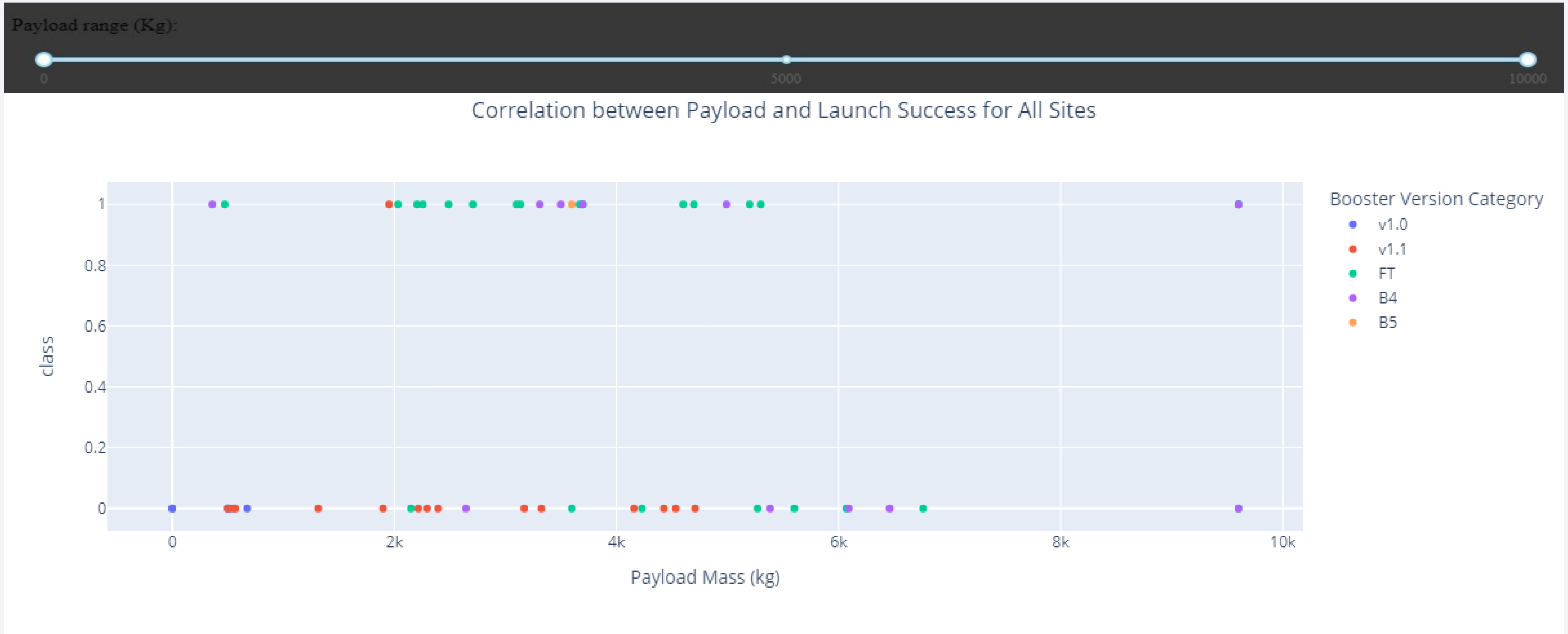
Highest Success outcome



KSC LC-39A site had the highest launch success rate of approximately 77%.

Payload vs. Launch Outcome

Scatter point of Payload vs. Launch outcome



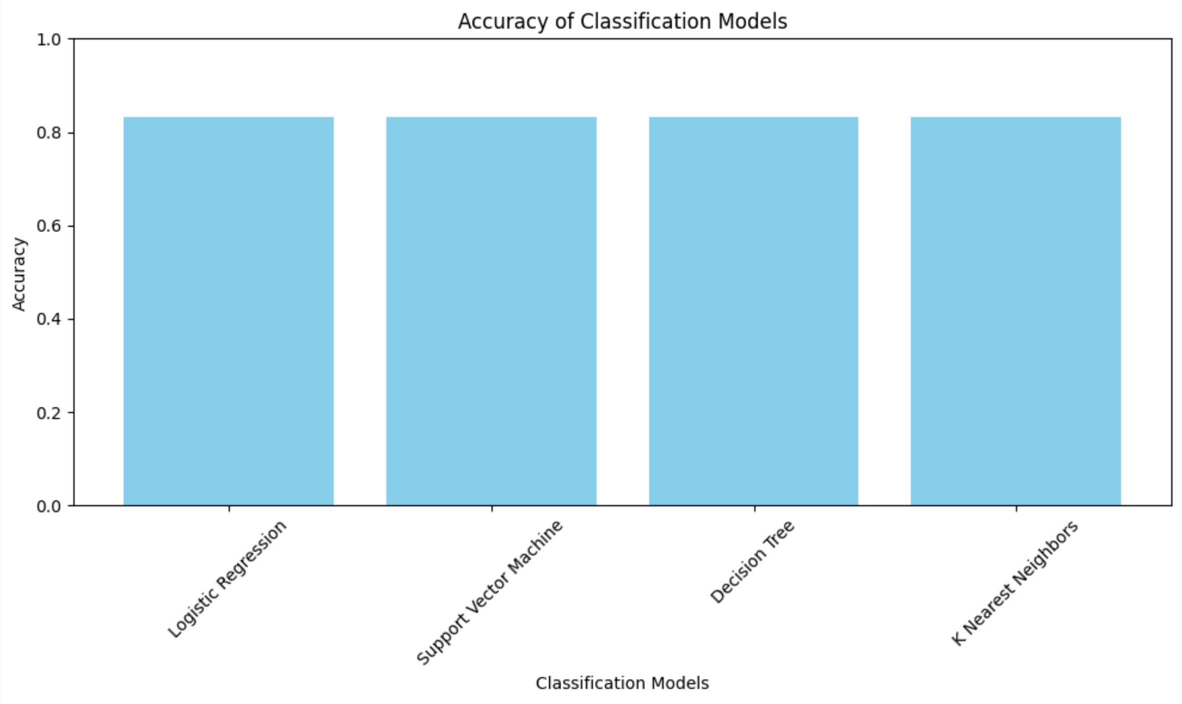
Payload range between 2,000 kg to 6,000 kg have the largest success rate.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

Bar chart of Accuracy for each Model on test data

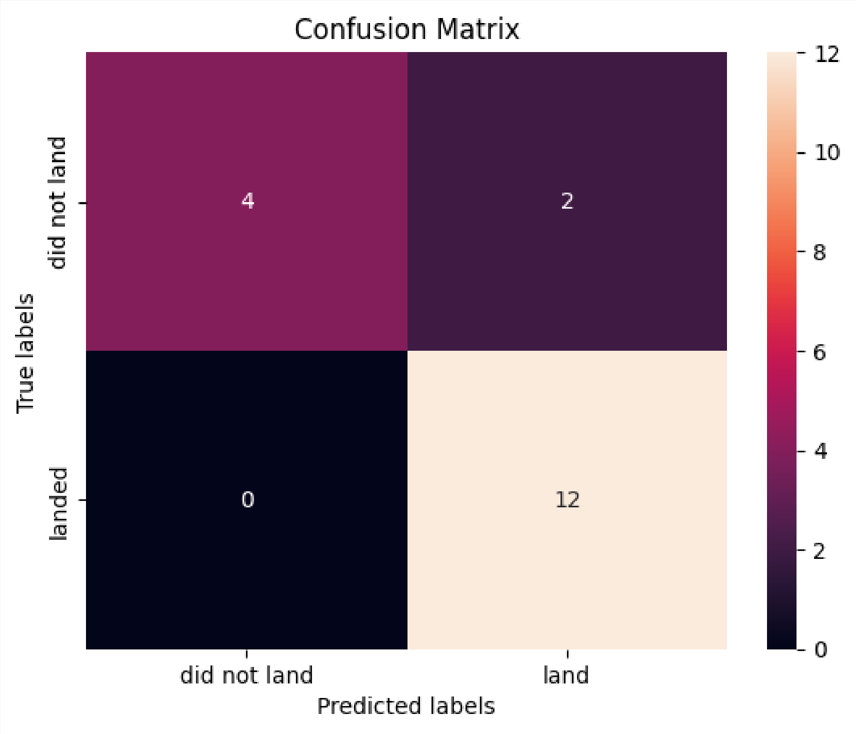


	Model	Precision	Recall	F1-score
0	Logistic Regression	0.8	1.0	0.888889
1	Support Vector Machine	0.8	1.0	0.888889
2	Decision Tree	0.8	1.0	0.888889
3	K Nearest Neighbors	0.8	1.0	0.888889

All four models have equally high classification accuracy, Precision, Recall and F1-score. However, I used the Decision Tree model because of its Confusion Matrix.

Confusion Matrix

Confusion matrix of Decision Tree model



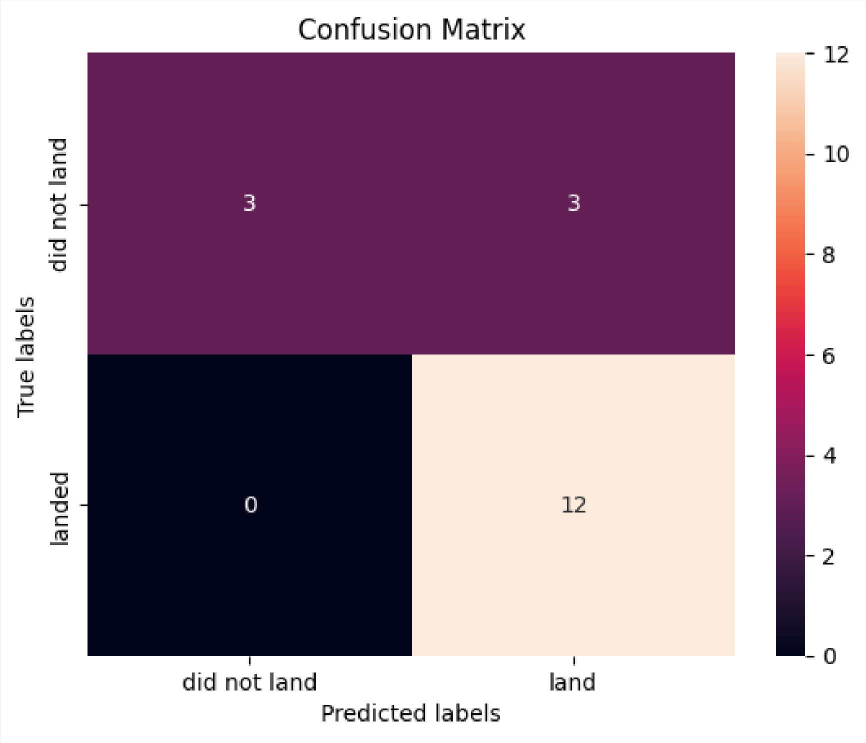
The Decision Tree Confusion Matrix shows the model has high accuracy in classifying failed outcomes, i.e 4 out of 6, compared with the other models that classified 3 out of 6. This model gives higher accuracy in predicting True Negatives (TN).

Conclusions

- The EDA revealed insights into the SpaceX Falcon9 dataset, showing how different factors interact.
- With an Interactive analytics and dashboard, we visualized the geographic locations of each launch site and their outcomes.
- We used rocket launch factors to build analytic models to predict landing outcomes for the first stage and a Decision Tree model proved to give high accuracy and classification of outcomes.
- Based on the Decision Tree's accuracy and Confusion Matrix, the predicted price for a SpaceY rocket launch could be estimated around $62 million, aligning closely with SpaceX's advertised cost. However, further analysis and validation against additional data sources are advisable to refine and adjust the pricing model for SpaceY's specific operational and cost structures.
- The Decision Tree model predicts with 83.3% accuracy that SpaceX will reuse the first-stage rocket for its launch. The Confusion Matrix further supports this prediction with 12 True positives and only 2 False Negatives, indicating a higher likelihood that SpaceX will indeed reuse the first stage.

Appendix

Confusion Matrix of Logistic Regression, Support Vector Matrix and K Nearest Neighbor



Appendix

GitHub repo:

https://github.com/TobiKehinde/IBM_Applied-Data-Science-Capstone.git

Thank you!