IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Prasun Maltare
26-2-2024

Outline

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

Executive Summary

In this capstone project, the aim is to forecast the success of the SpaceX Falcon 9 first stage landing through the application of various machine learning classification algorithms. The primary stages of this project involve:

- Gathering, organizing, and refining data
- Conducting exploratory data analysis
- Creating interactive data visualizations
- Implementing machine learning models for prediction The analysis reveals that certain attributes of the rocket launches exhibit a correlation with the launch outcomes, whether they are successful or unsuccessful. Furthermore, it is inferred that the decision tree algorithm stands out as potentially the most effective approach for predicting the success of Falcon 9 first stage landings.

Introduction

This capstone project aims to predict the successful landing of the Falcon 9 first stage. SpaceX advertises Falcon 9 launches at $62 million, significantly lower than competitors due to first stage reuse. Understanding first stage landing success is crucial for cost estimation and competitive bidding. Unsuccessful landings, often intentional, occur occasionally. The primary question is: Can we predict first stage landing success based on launch features like payload mass, orbit type, and launch site?

Section 1

Methodology

Methodology

Executive Summary

Data collection, wrangling, and formatting, using:

- SpaceX API
- Web scraping

Exploratory data analysis (EDA), using:

- Pandas and NumPy
- SQL
- Data visualization, using:

Matplotlib and Seaborn

- Folium
- Dash
- Machine learning prediction, using
- Logistic regression
- Support vector machine (SVM)

Data Collection

SpaceX API

The API used is https://api.spacexdata.com/v4/rockets/.

The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.

Every missing value in the data is replaced the mean the column that the missing value belongs to.

Web scraping

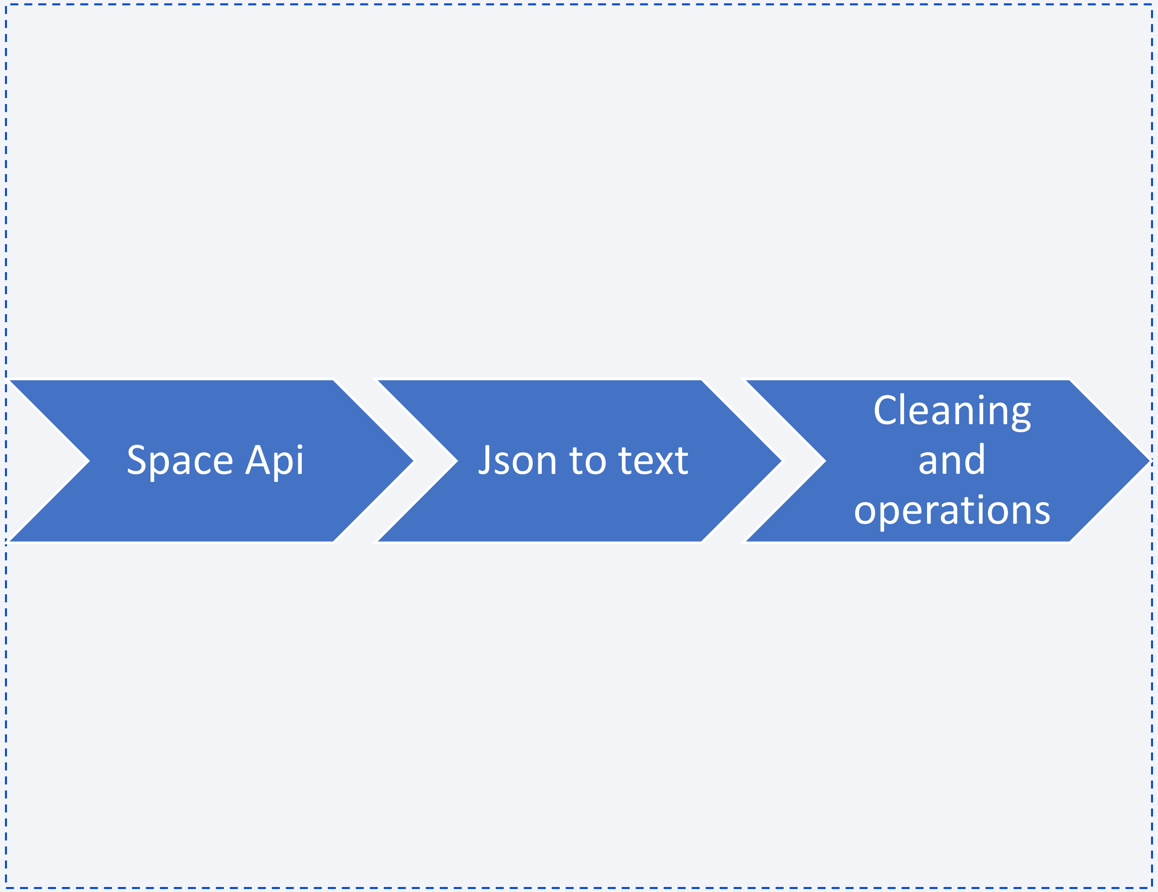
The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

The website contains only the data about Falcon 9 launches.

We end up with 121 rows or instances and 11 columns or features.

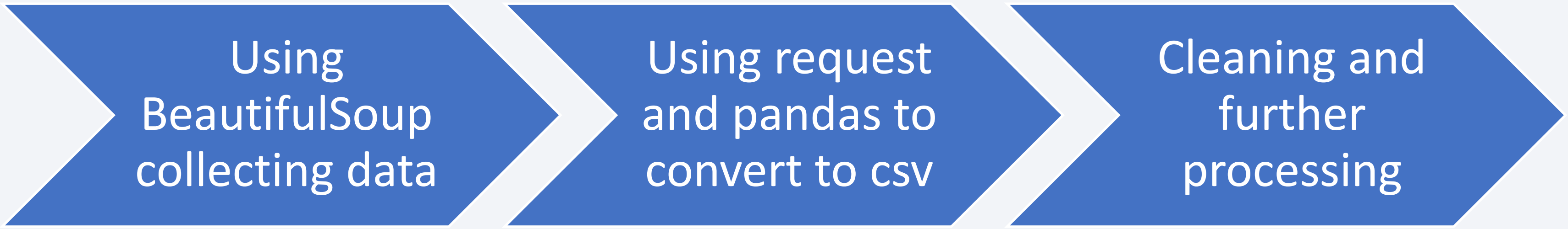
Data Collection – SpaceX API

- Github:https://github.com/Prasun157
- Using this api and cleaning data and performing EDA
- Check github using link for more info



Data Collection - Scraping

- Github URL



Data Wrangling

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.

- Github url:https://github.com/Prasun157/Final-capstone-project-data-science/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Using Pandas and NumPy:

- Counting launches per launch site.
- Tracking orbit occurrences.
- Analyzing mission outcome frequencies.

.

- Github url: https://github.com/Prasun157/Final-capstone-project-data-science/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

With SQL:

- Retrieving unique launch site names.
- Calculating total payload mass for NASA (CRS) boosters.
- Computing average payload mass for booster version F9 v1.1

- Github url https://github.com/Prasun157/Final-capstone-project-data-science/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- Matplotlib and Seaborn libraries are employed for data visualization, showcasing relationships between various features using scatterplots, bar charts, and line charts. Key visualizations include:
- - Exploring the correlation between flight number and launch site.
- - Analyzing the relationship between payload mass and launch site.
- - Investigating the connection between success rate and orbit type.
- Folium library is utilized for interactive mapping, enabling:
- - Plotting all launch sites on a map.
- - Distinguishing between successful and failed launches for each site.
- - Displaying distances from launch sites to nearby landmarks like cities, railways, or highways.
- https://github.com/Prasun157/Final-capstone-project-data-science/blob/main/jupyter-labs-Interactive-Visual-Analytics-with-Folium.ipynb

Build a Dashboard with Plotly Dash

- Dash
  - Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
  - Using a pie chart and a scatterplot, the interactive site shows:
    - The total success launches from each launch site
    - The correlation between payload mass and mission outcome (success or failure) for each launch site
- url: https://github.com/Prasun157/Final-capstone-project-data-science/blob/main/python%20dashly%20final.py

Predictive Analysis (Classification)

- The Scikit-learn library functions are employed to develop machine learning models. The prediction phase encompasses these steps:

- 1. Standardizing the data.
- 2. Splitting the data into training and test sets.
- 3. Creating machine learning models, including:
-    - Logistic regression
-    - Support vector machine (SVM)
-    - Decision tree
-    - K nearest neighbors (KNN)
- 4. Fitting the models to the training set.
- 5. Identifying the optimal combination of hyperparameters for each model.
- 6. Evaluating the models based on accuracy scores and confusion matrices.
- url:https://github.com/Prasun157/Final-capstone-project-data-science/blob/main/jupyter-labs-Machine-Learning-Prediction.ipynb

Results

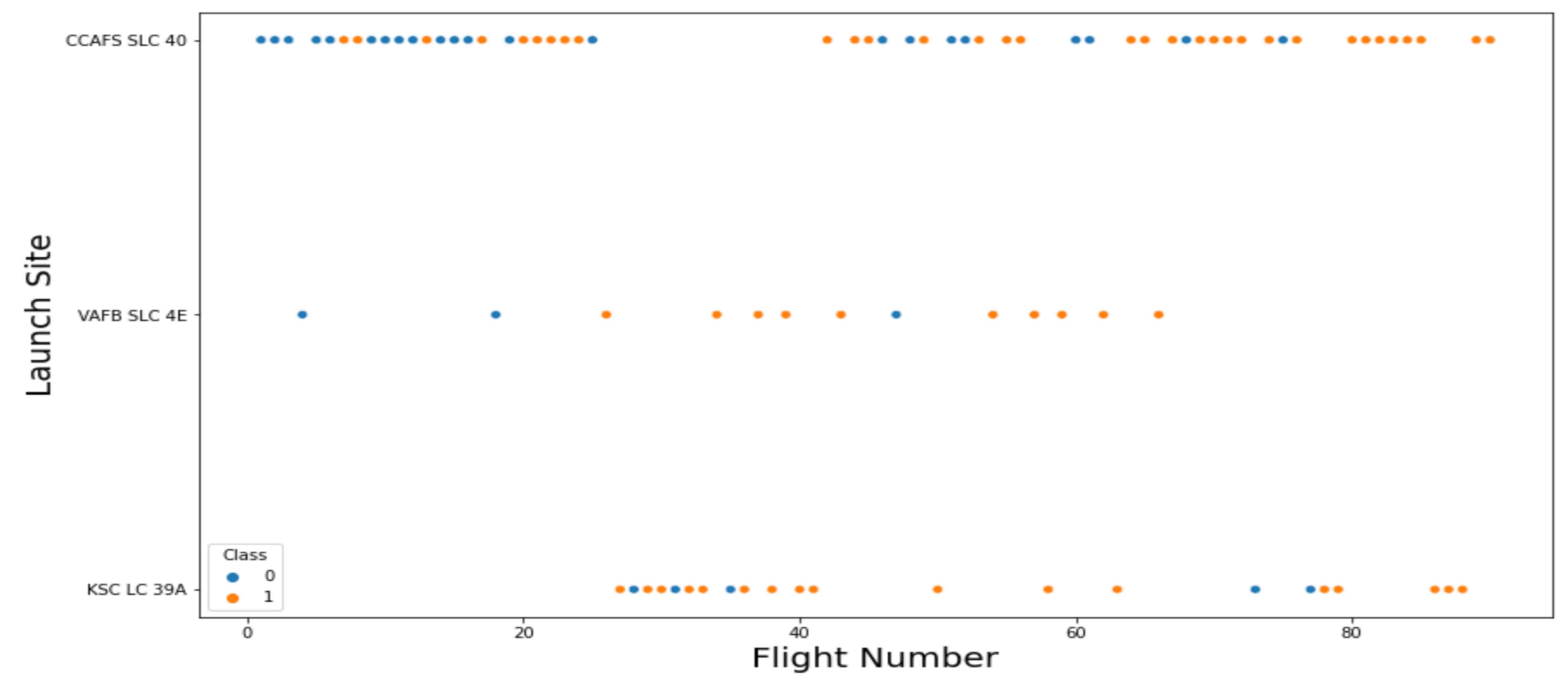
- Exploratory data analysis results
    - Both API and web scraping are capable to collect Xspace data
- Interactive analytics demo in screenshots
    - EDA with SQL is effective for data filtering
    - EDA with interactive visualization provides informative information
    - Plotly Dash is powerful to show instant data change
- Predictive analysis results
    - Decision Tree Classifier Algorithm has the best accuracy of predicting.

Section 2

Insights drawn from EDA

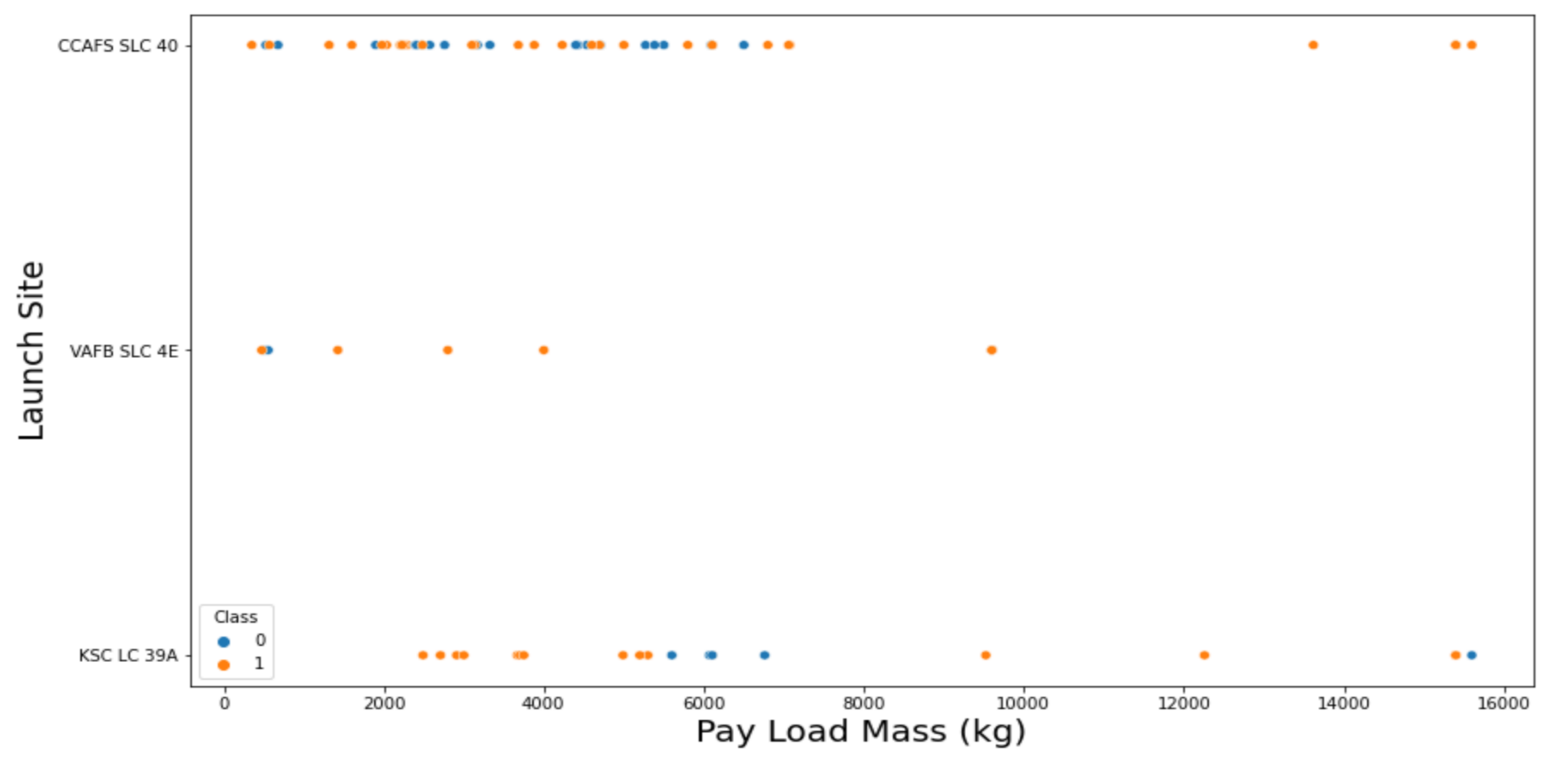
Flight Number vs. Launch Site

- scatter plot of Flight Number vs. Launch Site



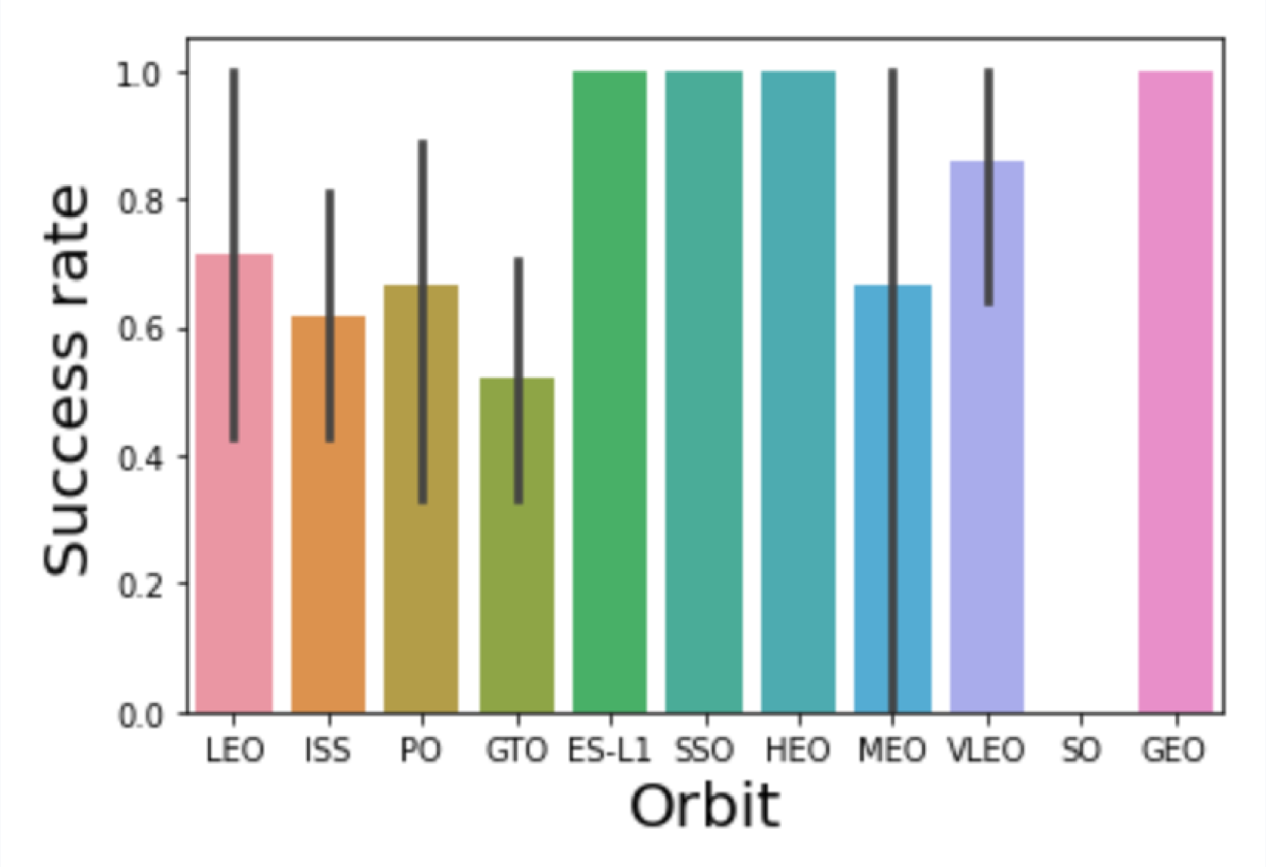
Payload vs. Launch Site

- scatter plot of Payload vs. Launch Site



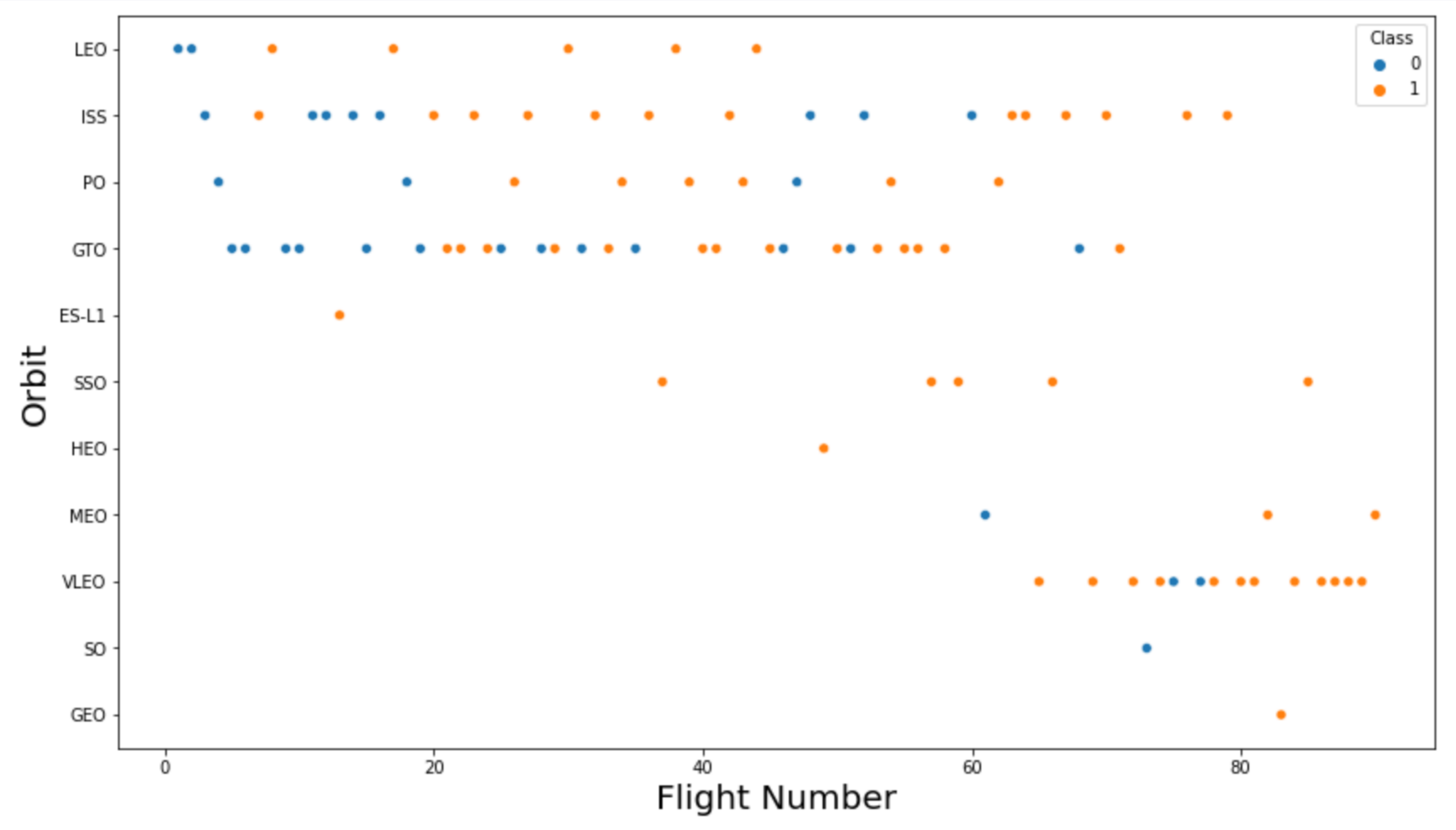
Success Rate vs. Orbit Type

- bar chart for the success rate of each orbit type



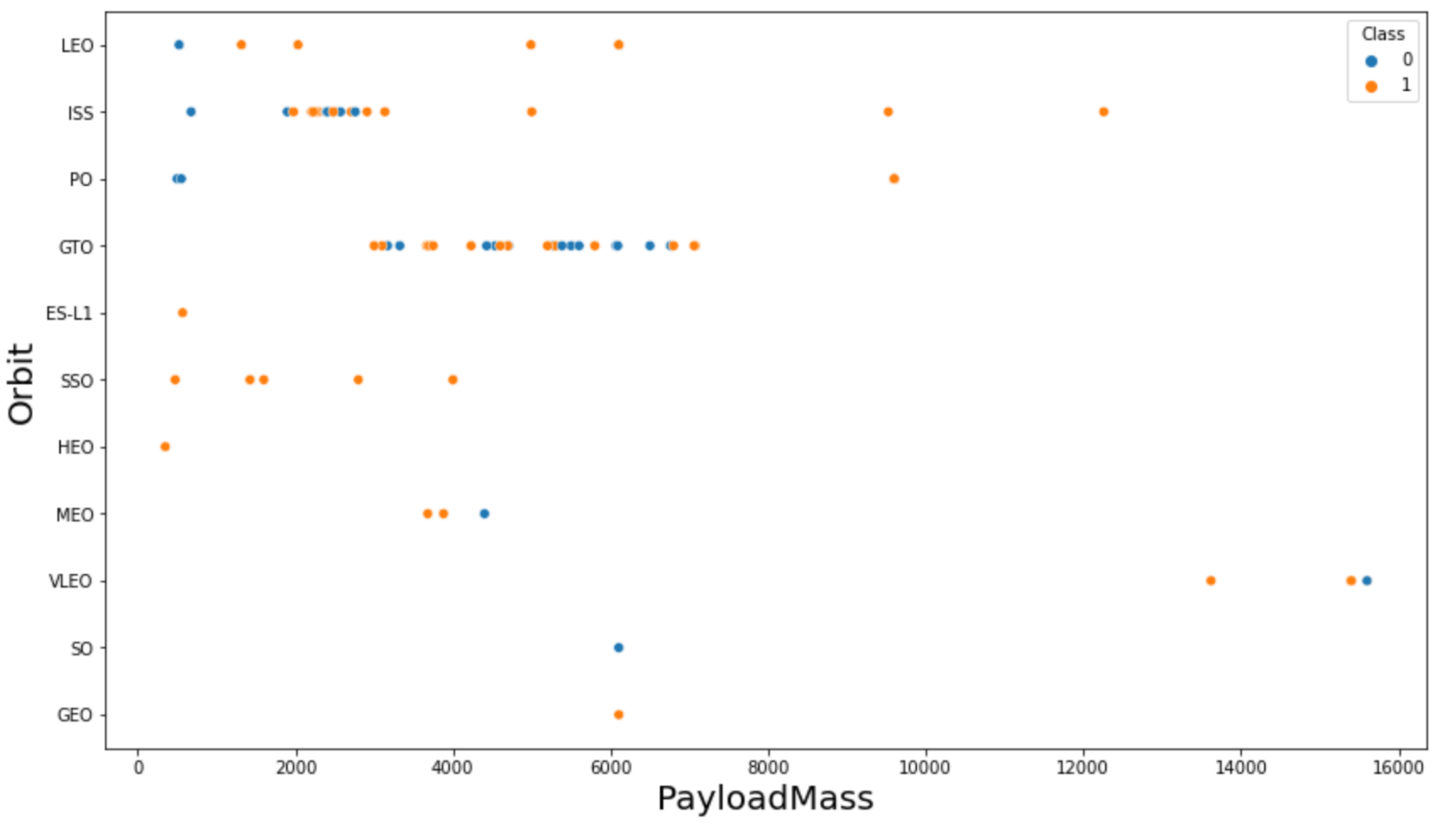
Flight Number vs. Orbit Type

- scatter point of Flight number vs. Orbit type



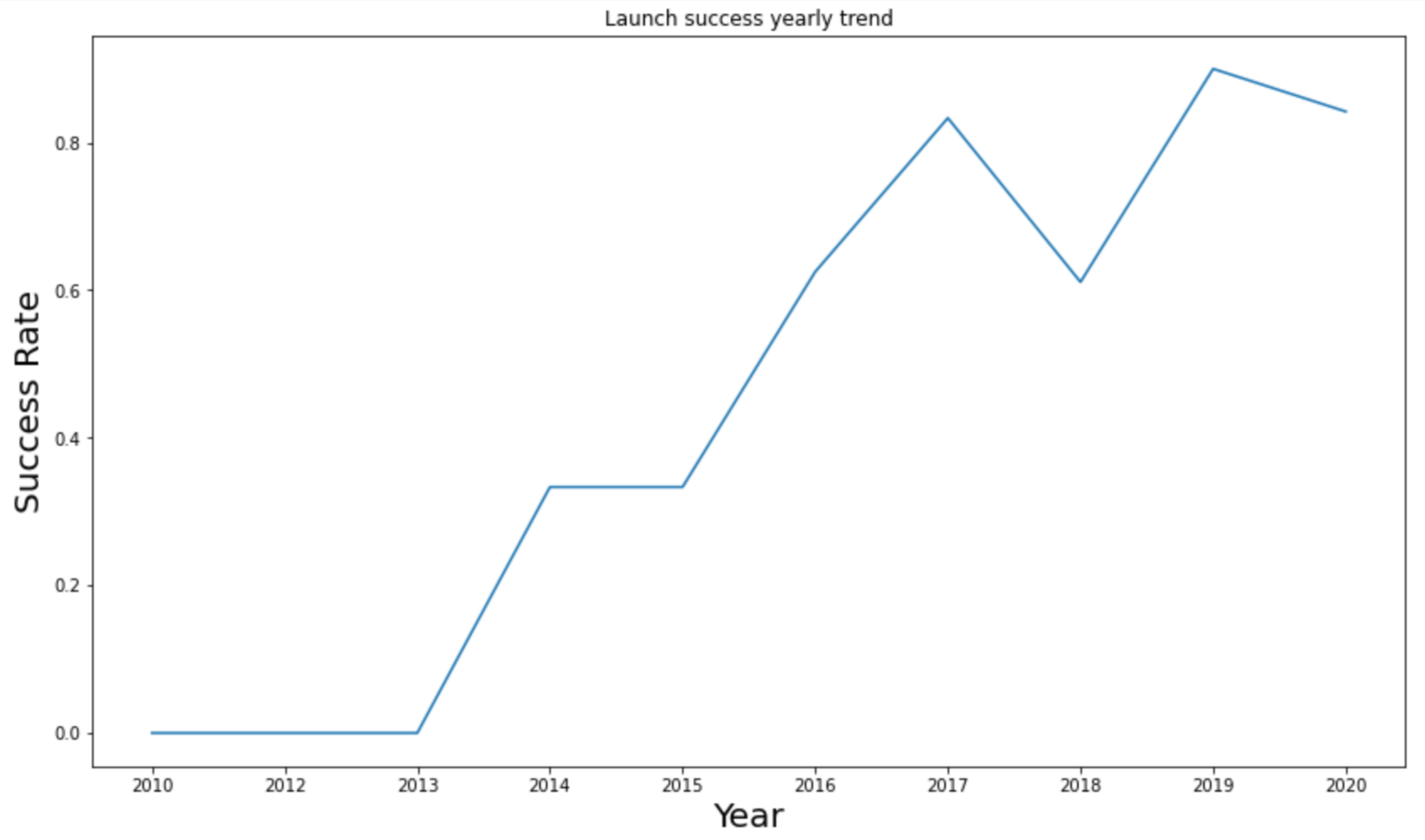
Payload vs. Orbit Type

- scatter point of payload vs. orbit type



Launch Success Yearly Trend

- line chart of yearly average success rate



All Launch Site Names

- the names of the unique launch sites
- Here is the name of all launch site , using sql queries

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

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	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 and Average Payload Mass by F9 v1.1 And First Successful Ground Landing Date

Total payload mass by NASA (CRS)
45596

Average payload mass by Booster Version F9 v1.1
2928

Date of first successful landing outcome in ground pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the 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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

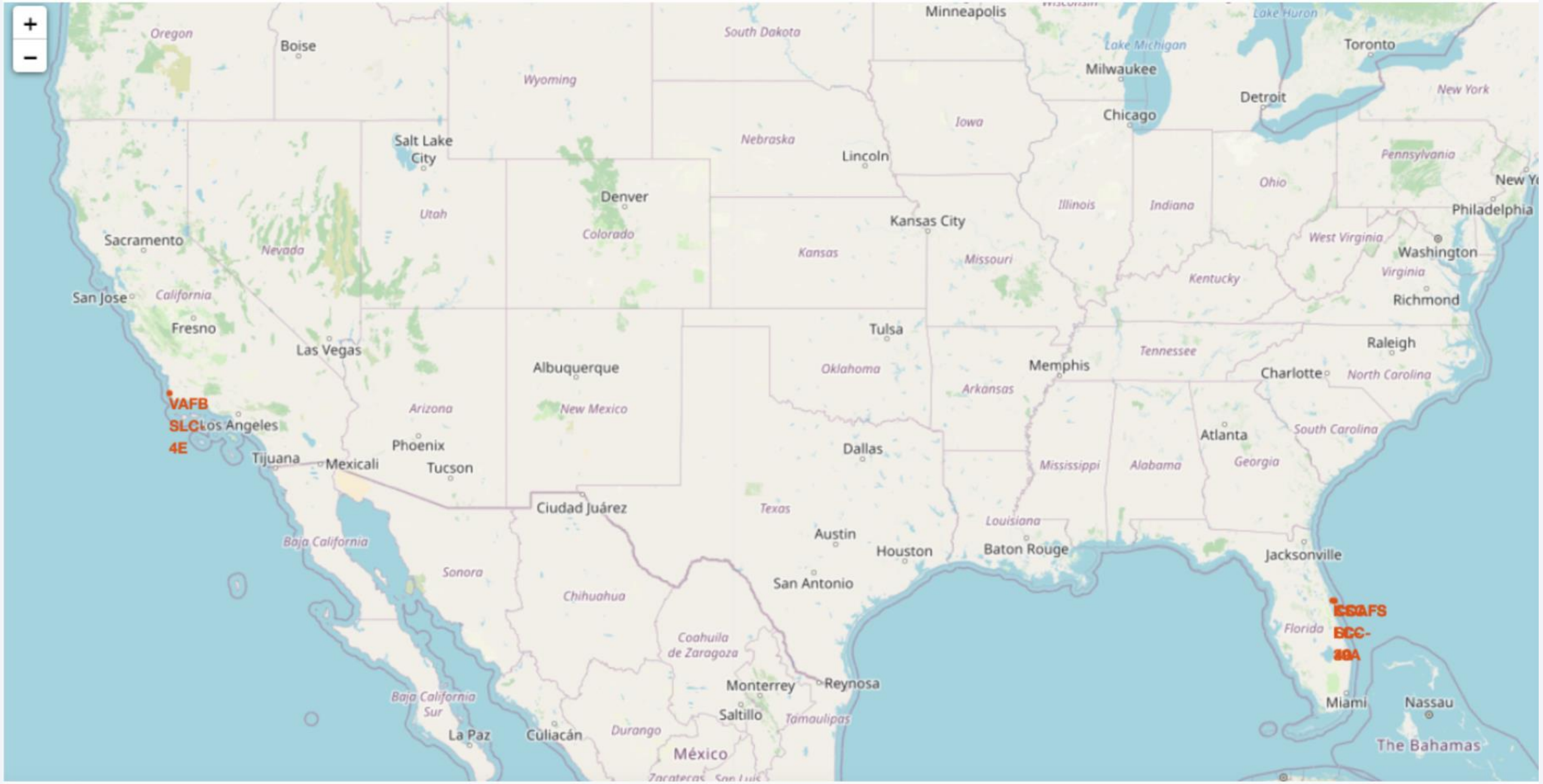
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

landing__outcome	landing_count
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

Section 3

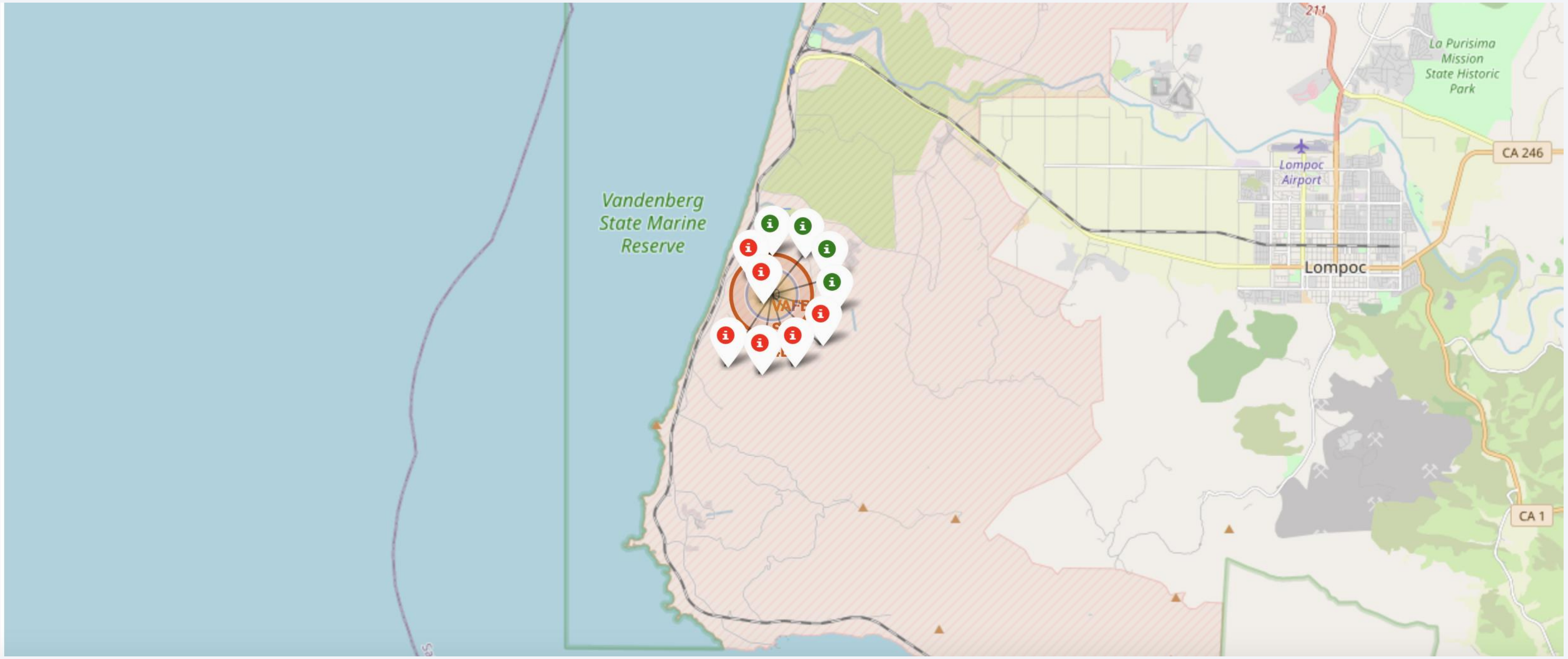
Launch Sites Proximities Analysis

All launch sites on map

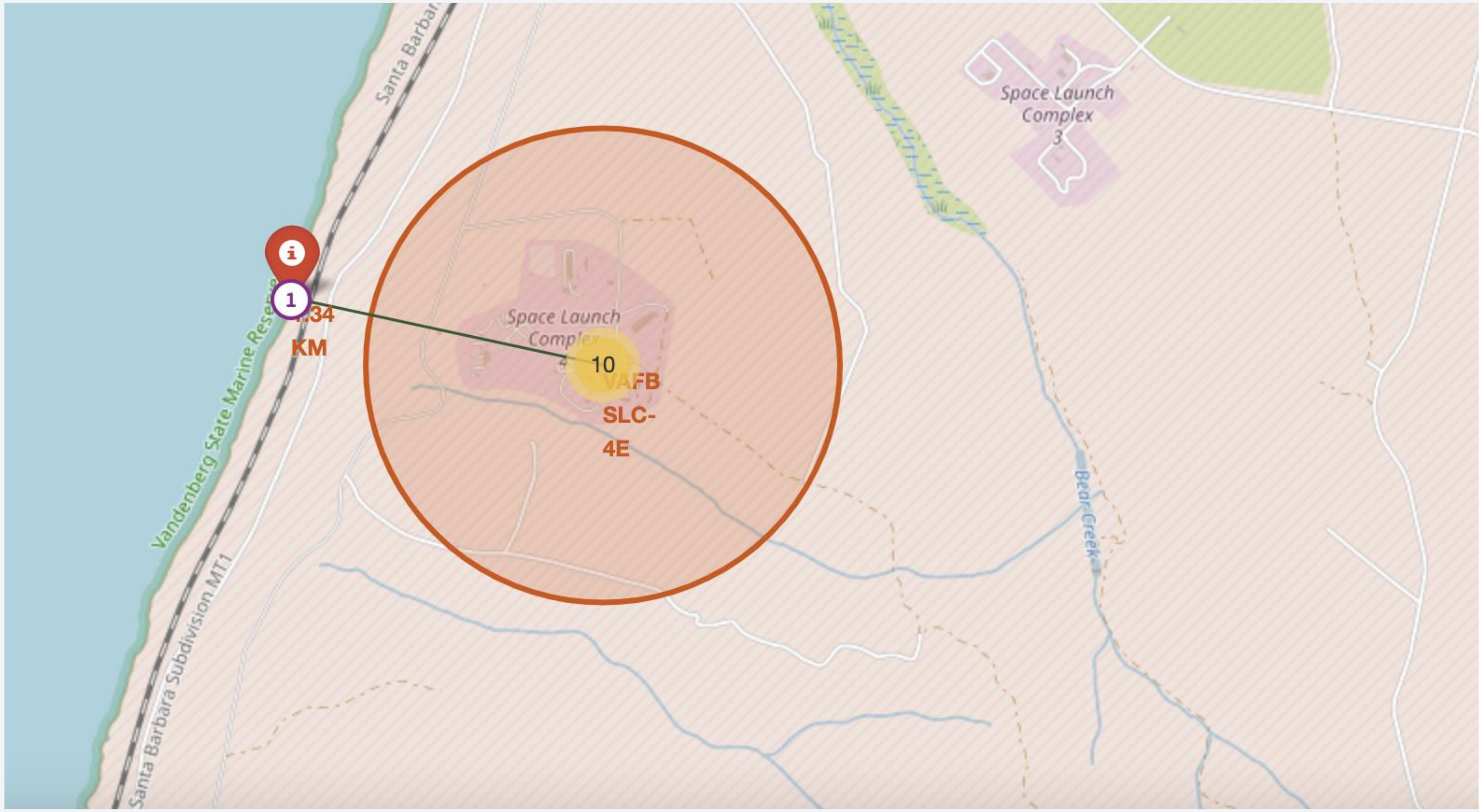


The succeeded launches and failed launches for each site on map

If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch



The distances between a launch site to its proximities such as the nearest city, railway, or highway

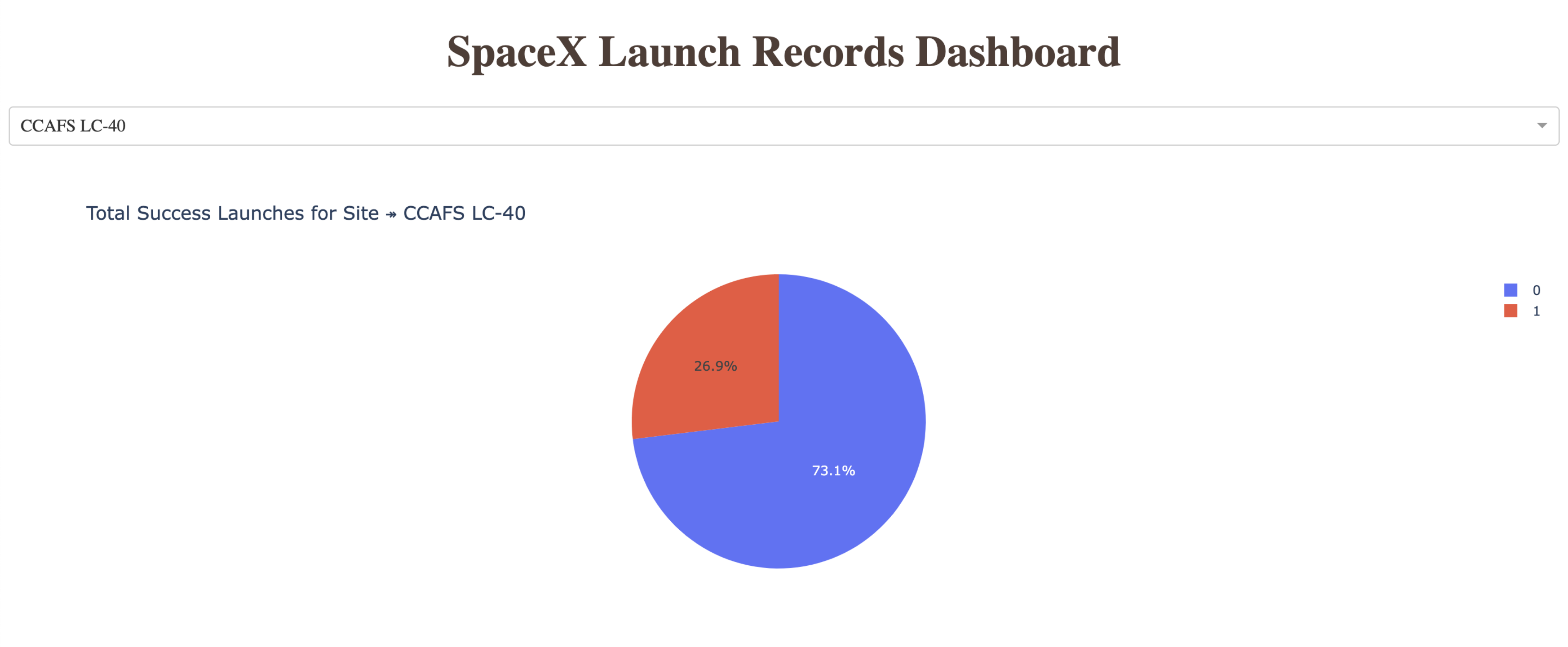


Section 4

Build a Dashboard with Plotly Dash

a pie chart when launch site CCAFS LC-40

The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.



scatterplot when the payload mass range is set to be from 2000kg to 8000kg

The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg. Class 0 represents failed launches while class 1 represents successful launches.

Section 5

Predictive Analysis (Classification)

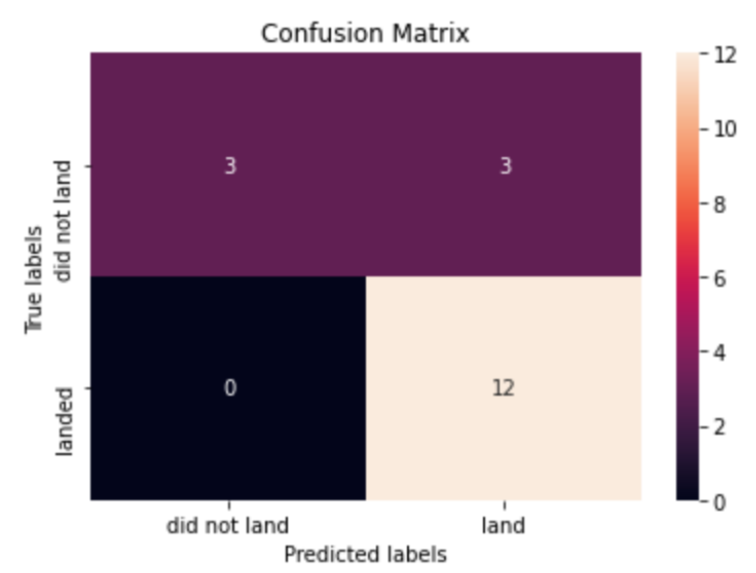
Classification Accuracy

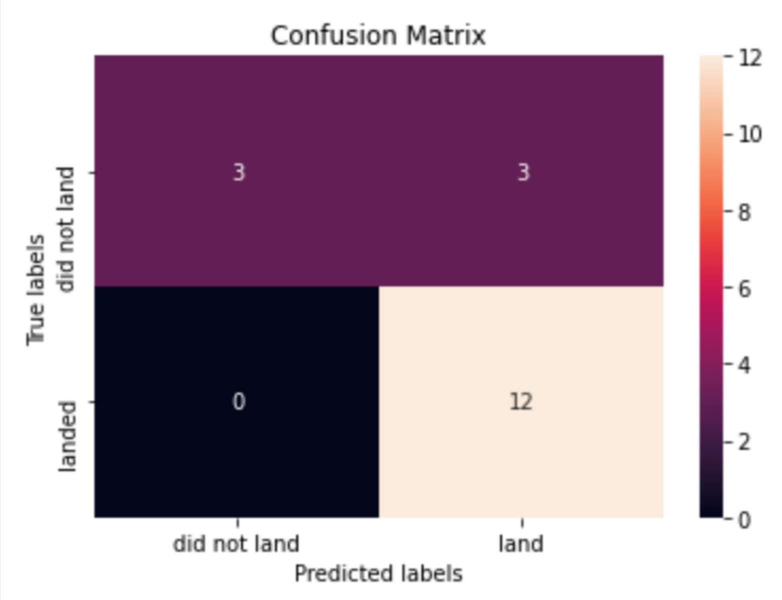
- Based on the GridSearchCV best scores, the models are ranked as follows, with the first being the best and the last being the worst:

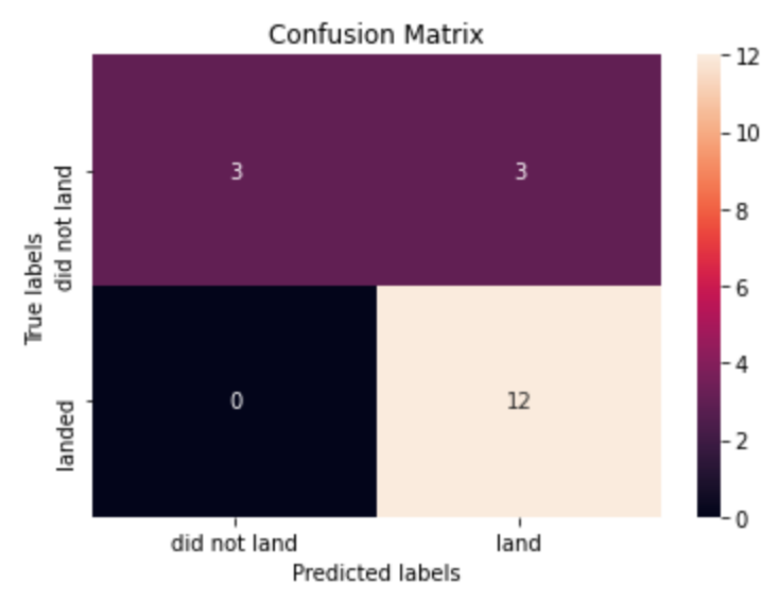
- 1. Decision tree (GridSearchCV best score: 0.8892857142857142)
- 2. K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
- 3. Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
- 4. Logistic regression (GridSearchCV best score: 0.8464285714285713)

Confusion Matrix

- 1. Decision tree (GridSearchCV best score: 0.8892857142857142)
- 2. K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
- 3. Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
- 4. Logistic regression (GridSearchCV best score: 0.8464285714285713)







Conclusions

In this project, the goal is to predict the landing success of the first stage in Falcon 9 launches to estimate launch costs accurately. Various features of a Falcon 9 launch, like payload mass and orbit type, may influence the mission outcome differently.

Multiple machine learning algorithms are utilized to analyze historical Falcon 9 launch data and generate predictive models. Among the four algorithms tested, the decision tree algorithm demonstrated the highest performance in predicting launch outcomes.

Appendix

- You can go through my git for all the codes and queries

- [my github]

Made by-
Prasun Maltare

Thank you!