IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Linnet Gomes
June 2024

Outline

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

Executive Summary

Process summary:

- Data collection
- Data wrangling
- Exploratory Data Analysis (EDA) with data visualization
- EDA with SQL
- Building an interactive map with Folium
- Building a Dashboard with Plotly Dash
- Predictive analysis – Classification

Results summary:

- EDA results
- Interactive results
- Predictive analysis

Introduction

Problem Statement:

- SpaceX advertises rocket launches on its website, with a cost of 62 million dollars, other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. If we can determine whether the first stage can land successfully then we can determine its cost of launch.

Problem Solution:

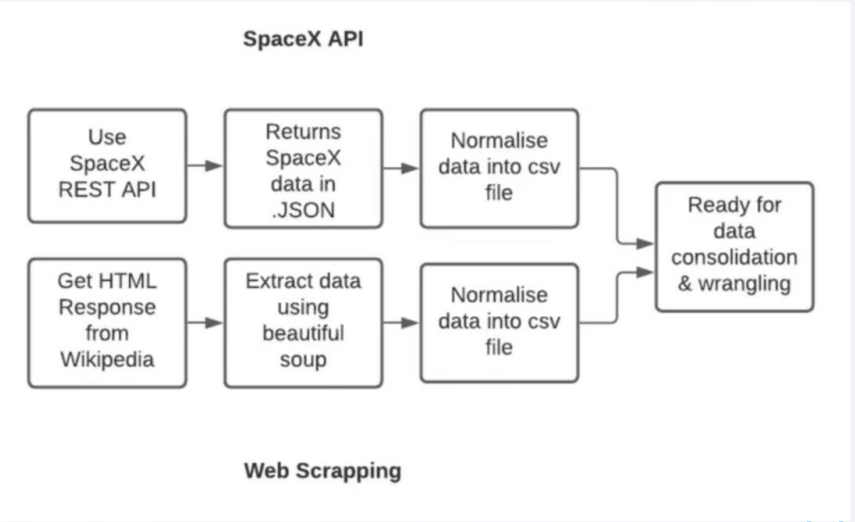
- This project is to predict if the Falcon 9 first stage will land successfully. Set of features are given about Falcon 9 rocket launch such as its payload mass, orbit type, launch site etc.

Section 1

Methodology

Methodology: Data Collection

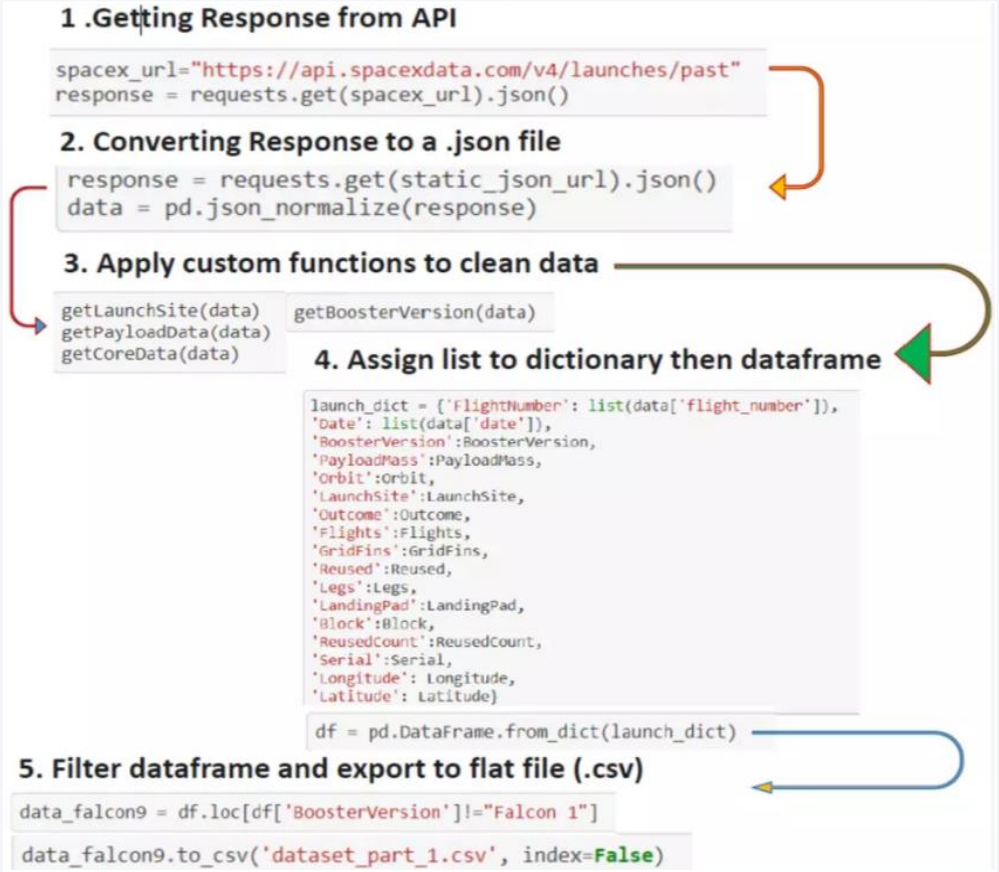
- SpaceX launch data that is gathered from REST API.
- This API will give us the required data about launches and rocket's set of features.
- The SpaceX REST API endpoints, or URL starts with api.spacexdata.com/v4/.
- Another popular method for obtaining required data is Web scraping Wikipedia using BeautifulSoup.



Methodology: Data Collection – SpaceX API

- Data collection with SpaceX REST API calls.

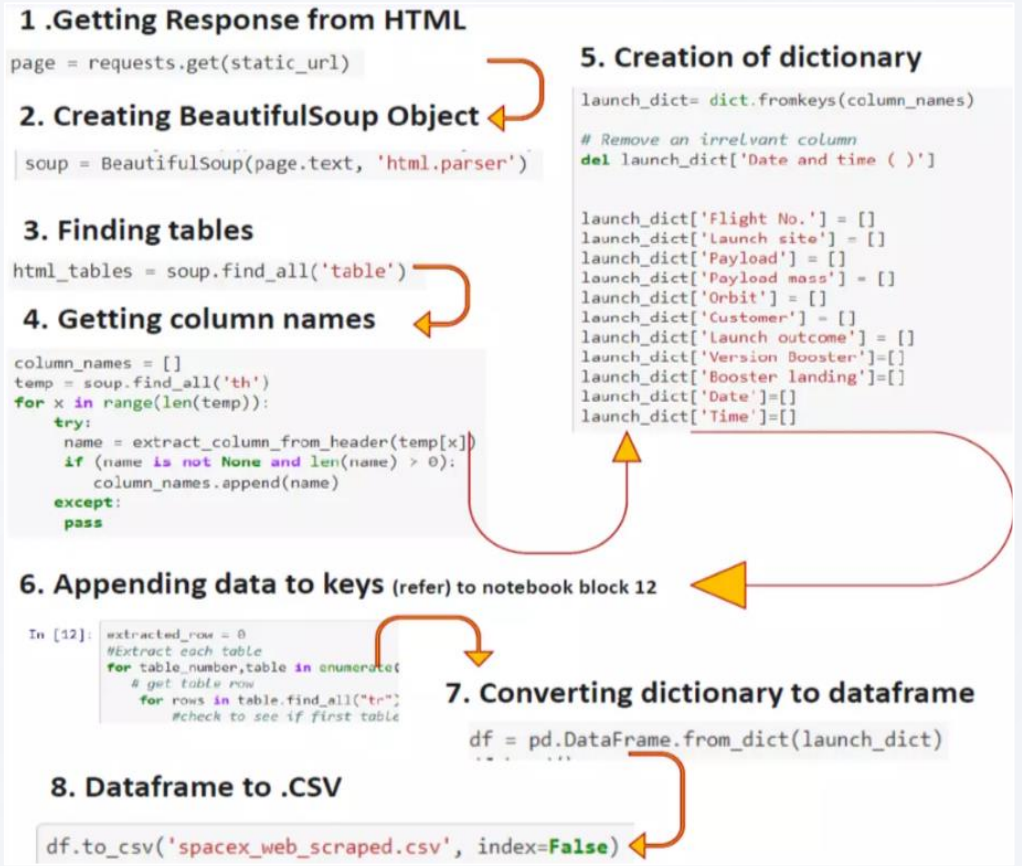
- https://github.com/LGLinnet/Final-Project/blob/main/spacex-data-collection-api.ipynb



Methodology: Data Collection - WebScraping

- Data collection using Webscraping from Wikipedia

- https://github.com/LGLinnet/Final-Project/blob/main/spacex-data-collection-webscraping.ipynb

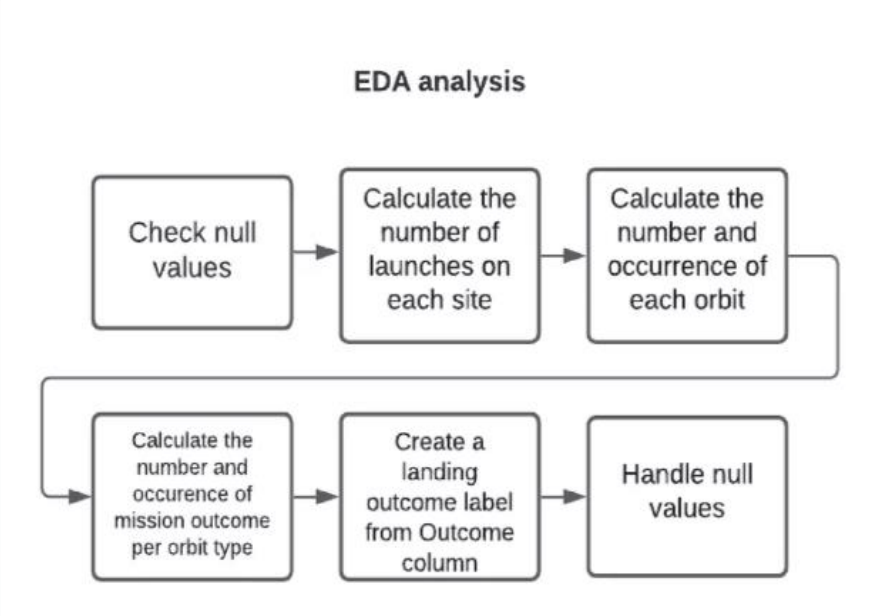


Methodology: Data Wrangling

- The data is later processed so that there are no missing entries and categorical values are encoded using one-hot encoding.

- An extra column called ‘Class’ is added to data frame. It contains 0 if launch is failed and 1 if successful.

- https://github.com/LGLinnet/Final-Project/blob/main/spacex-eda-Data%20wrangling.ipynb



Methodology: EDA with Data Visualization

- Pandas and NumPy:
  - Functions from Pandas and NumPy libraries are used to derive info. about data collected, which includes:
    - The no. of launches on each launch site
    - The no. and occurrence of each mission outcome
    - The no. of occurrence of each orbit

- https://github.com/LGLinnet/Final-Project/blob/main/spacex-eda-data%20visualization.ipynb

Methodology: EDA with SQL

- The data is queried using SQL to answer questions such as:
  - The names of unique launch sites
  - The total payload mass
  - The average payload mass

- https://github.com/LGLinnet/Final-Project/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Methodology: Build an Interactive Map with Folium

- Matplotlib and Seaborn:
  - Functions from Matplotlib and Seaborn libraries are used to visualize data through scatterplots, bar charts and line charts. They are used to understand about relation between features such as:
    - The relation b/w flight no. and launch site
    - The relation b/w payload mass and launch site
    - The relation b/w success rate and orbit type
- Folium
  - The data is queried using SQL to answer questions such as:
    - Mark all launch sites on map
    - Mark succeeded and failed launches for each site on map
    - Mark distance b/w launch site to its proximities such as nearest city, airport, coastline, highway etc.
- https://github.com/LGLinnet/Final-Project/blob/main/spacex-launchsite-analysis-Folium.ipynb

Methodology: Build a Dashboard with Plotly Dash

- Dash:
  - Functions from Dash are used to generate interactive site where input can be given using dropdown menu and range slider.
  - Using pie chart and scatterplot, the interactive site shows information such as:
    - The total successful launches from each launch site
    - The correlation b/w payload mass and mission outcome for each launch site
- https://github.com/LGLinnet/Final-Project/blob/main/spacex_dash_app.py

Methodology: Predictive Analysis (Classification)

- MACHINE LEARNING PREDICTION:
- Functions from Scikit library are used to create our machine learning models.
- The machine learning prediction phase includes:
  - Standardizing the data
  - Splitting data into training and test data
  - Creating machine learning models include:
    - Logistic regression
    - Support Vector Machine (SVM)
    - Decision tree
    - K nearest neighbours (KNN)
- Fit the models on training set
- Find the best combination of hyperparameters for each model
- Evaluate models based on their accuracy score and confusion matrix
- https://github.com/LGLinnet/Final-Project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

- EDA with SQL (SQL)
- EDA with Data Visualization (Matplotlib and Seaborn)
- Folium (Map)
- Dash (Dashboard)
- Predictive Analysis

Section 2

Insights drawn from EDA

RESULTS – EDA with SQL

1. Display the names of the unique launch sites in the space mission

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

2. Display 5 records where launch sites name starts 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

RESULTS – EDA with SQL

3. Display Total payload mass carried by boosters launched by NASA (CRS)

Payload mass
45596

4. Display avg payload mass carried by booster F9 v1.1

Average Payload mass
2534.6666666666665

5. Display date when first successful landing in ground pad achieved

First Date
2015-12-22

RESULTS – EDA with SQL

6. Display names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Successful Booster
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

7. Display total number of successful and failure mission outcomes

Mission Outcome	Outcome count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

RESULTS – EDA with SQL

8. Display the names of the booster versions which have carried the maximum payload mass

Max Payload Booster Name
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

9. Display the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in year 2015

Month	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

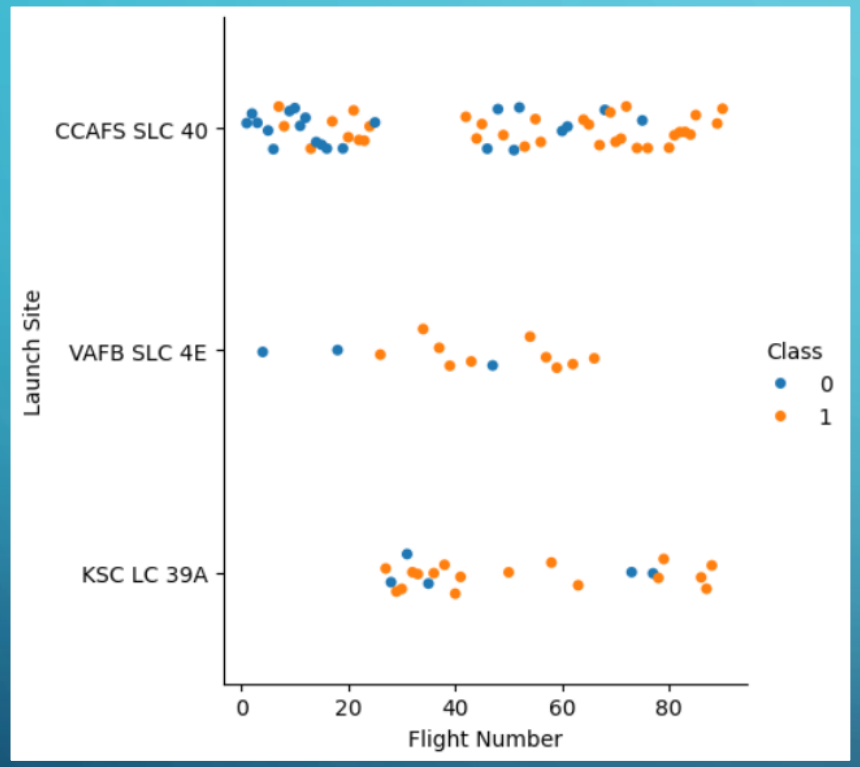
RESULTS – EDA with SQL

10. Display the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

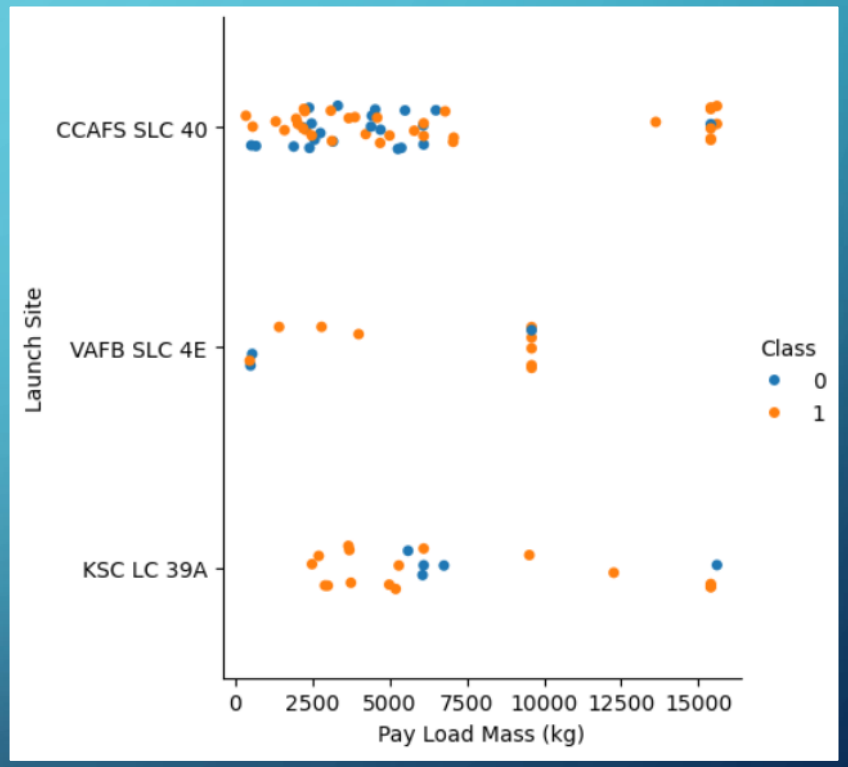
Landing Outcome	Outcome Count
No attempt	10
Success (drone ship)	6
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

RESULTS – EDA with Visualization

1. Relation b/w Flight no. and Launch site

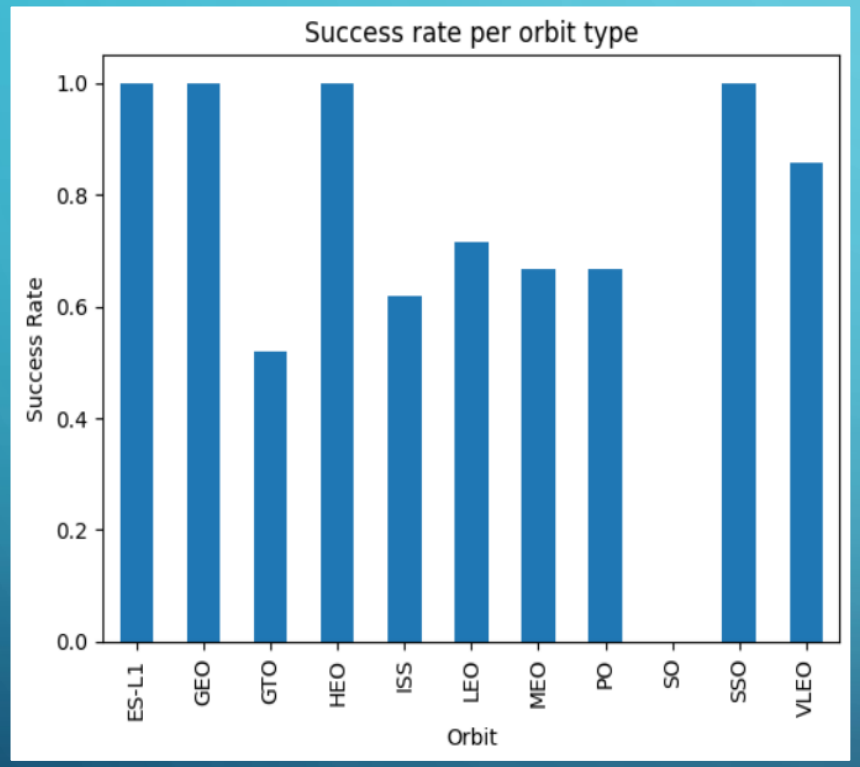


2. Relation b/w Payload mass and Launch site

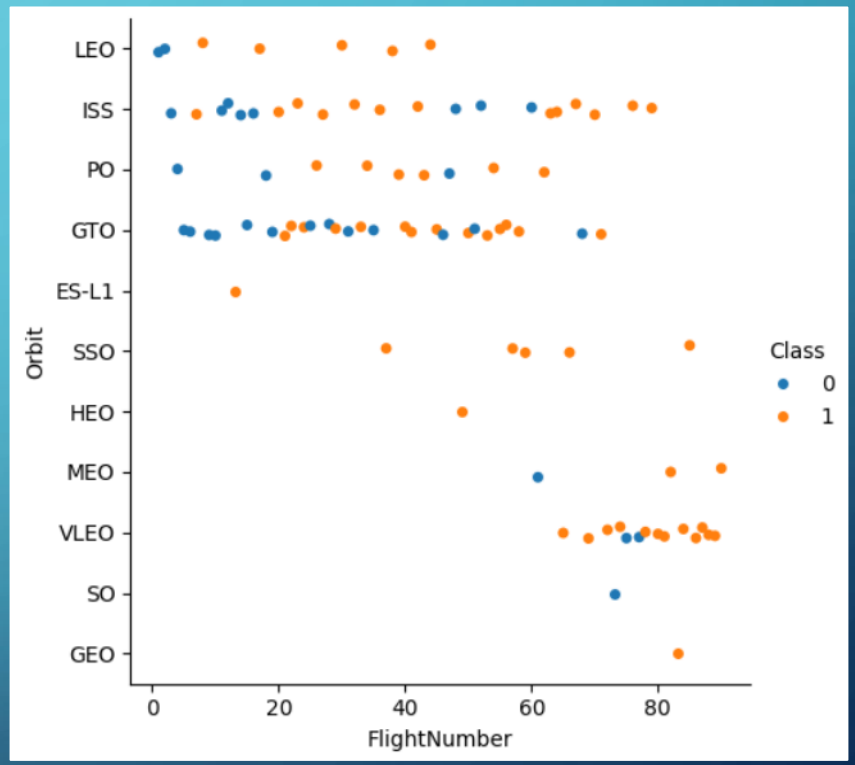


RESULTS – EDA with Visualization

3. Relation b/w Success rate and Orbit type

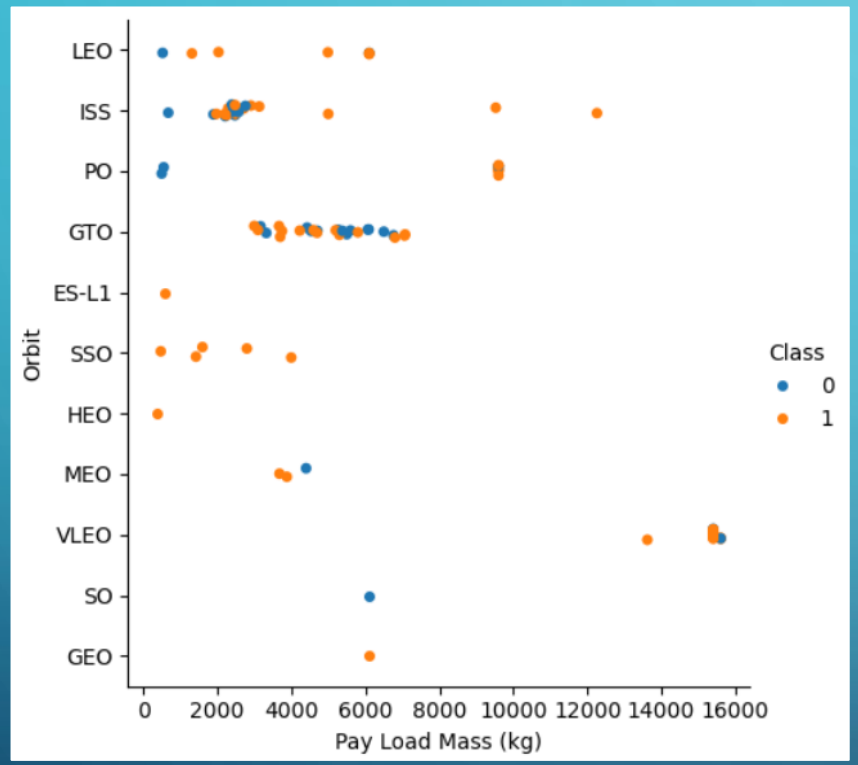


4. Relation b/w Flight no. and Orbit type

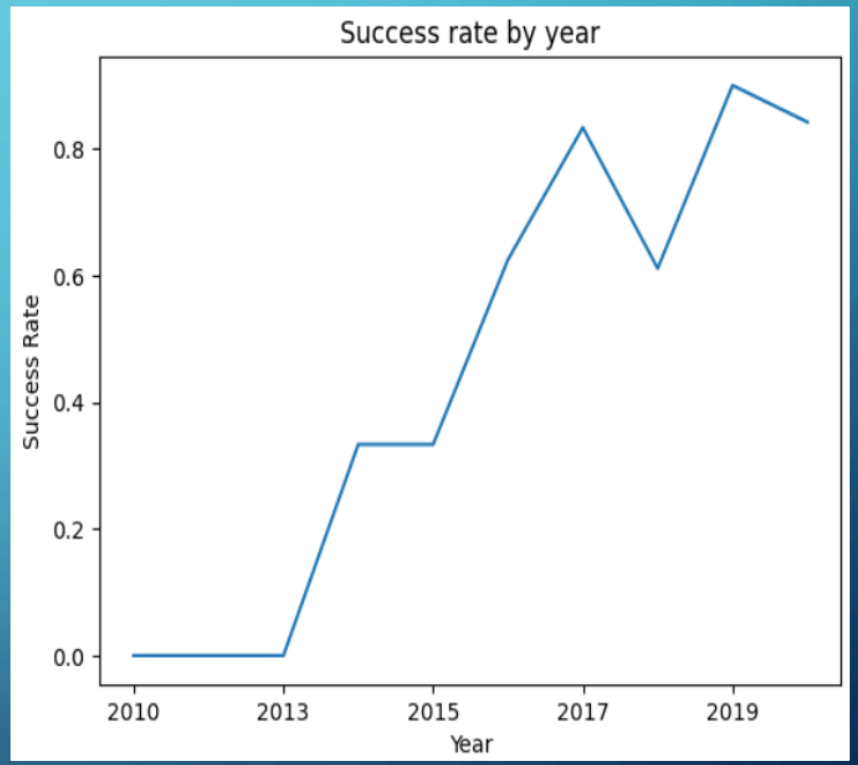


RESULTS – EDA with Visualization

5. Relation b/w Payload mass and Orbit type

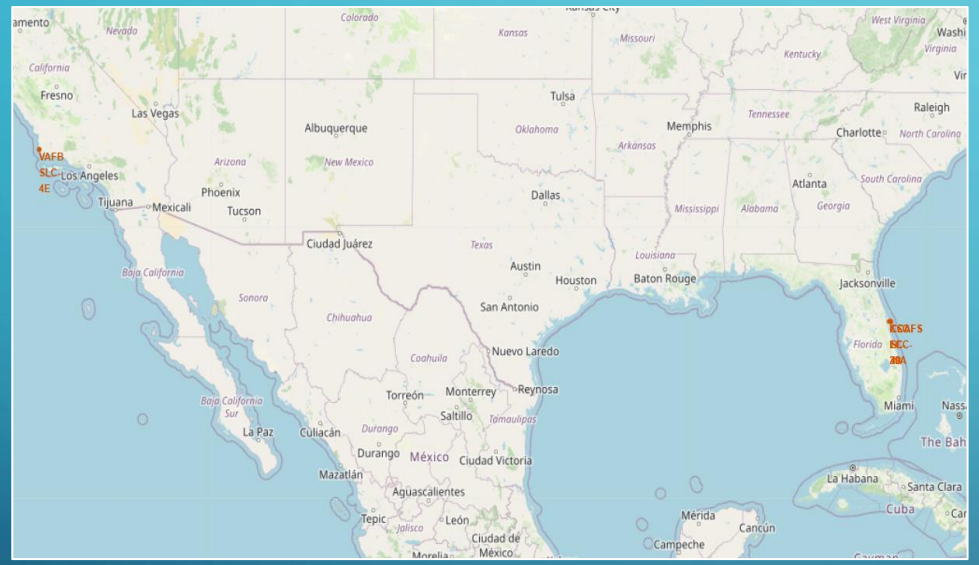


6. Launch success rate by year

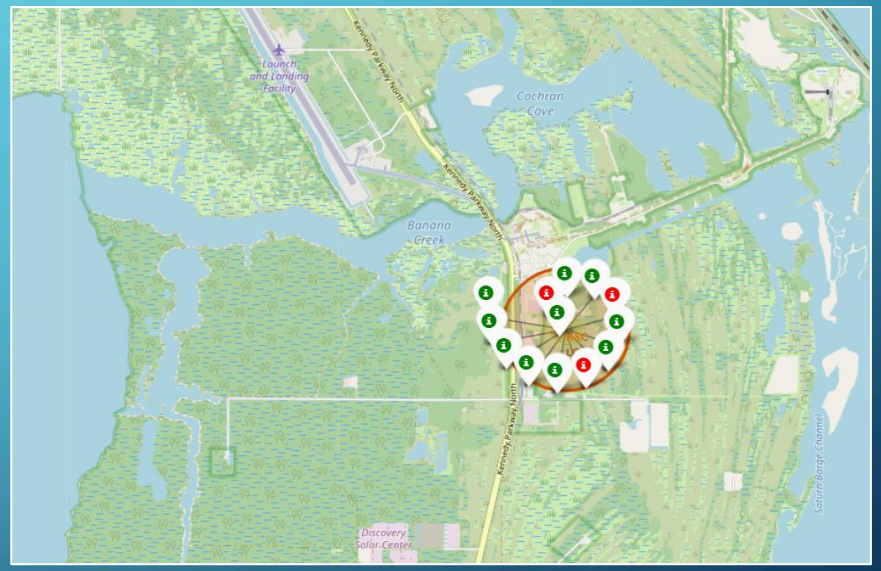


RESULTS – Folium

1. All Launch sites plotted on map

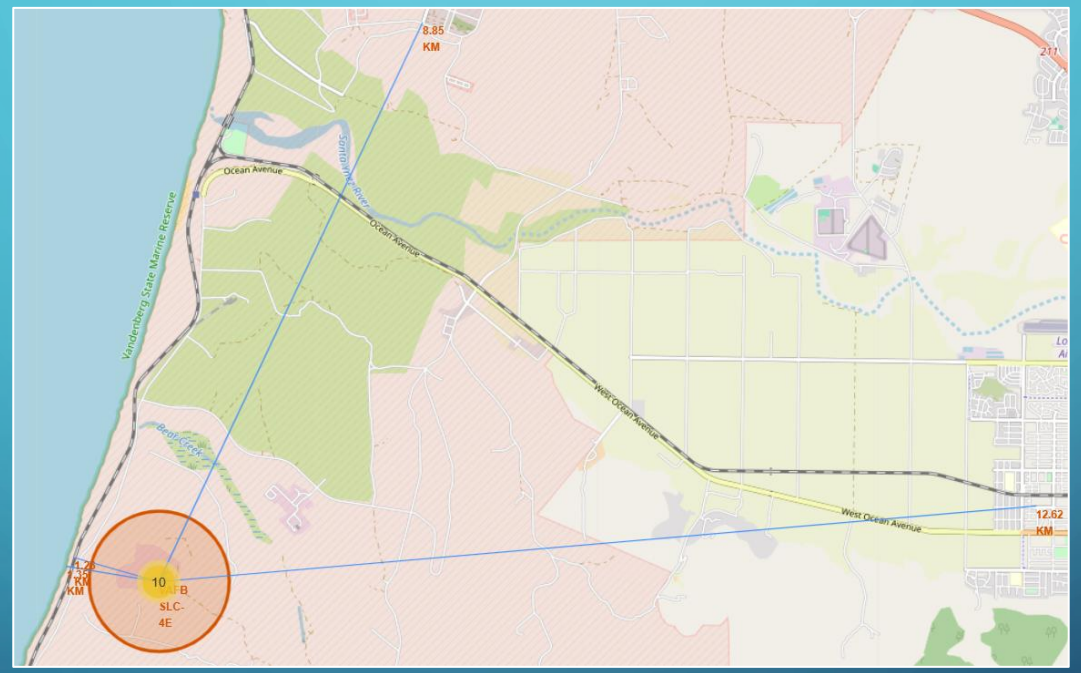


2. Successful and failed launches for each site on map (Green tag – success, Red tag – Fail)



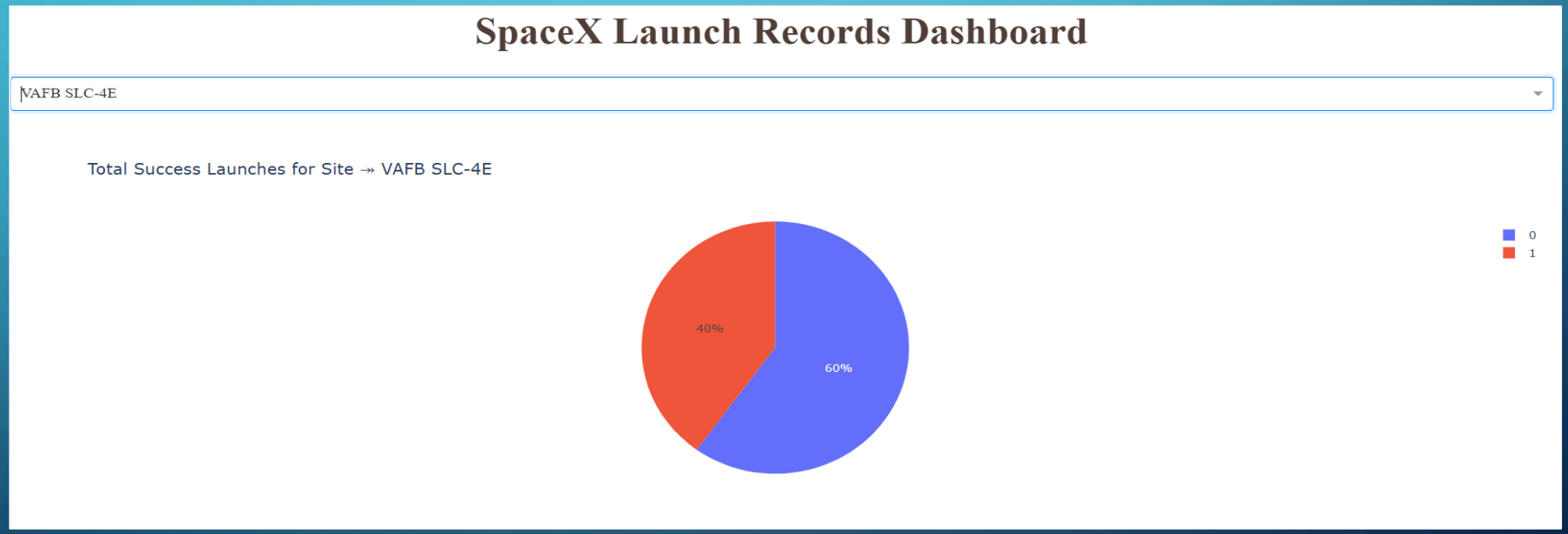
RESULTS – Folium

3. Distance between each launch sites and its proximities such as nearest city, airport, coastline etc.



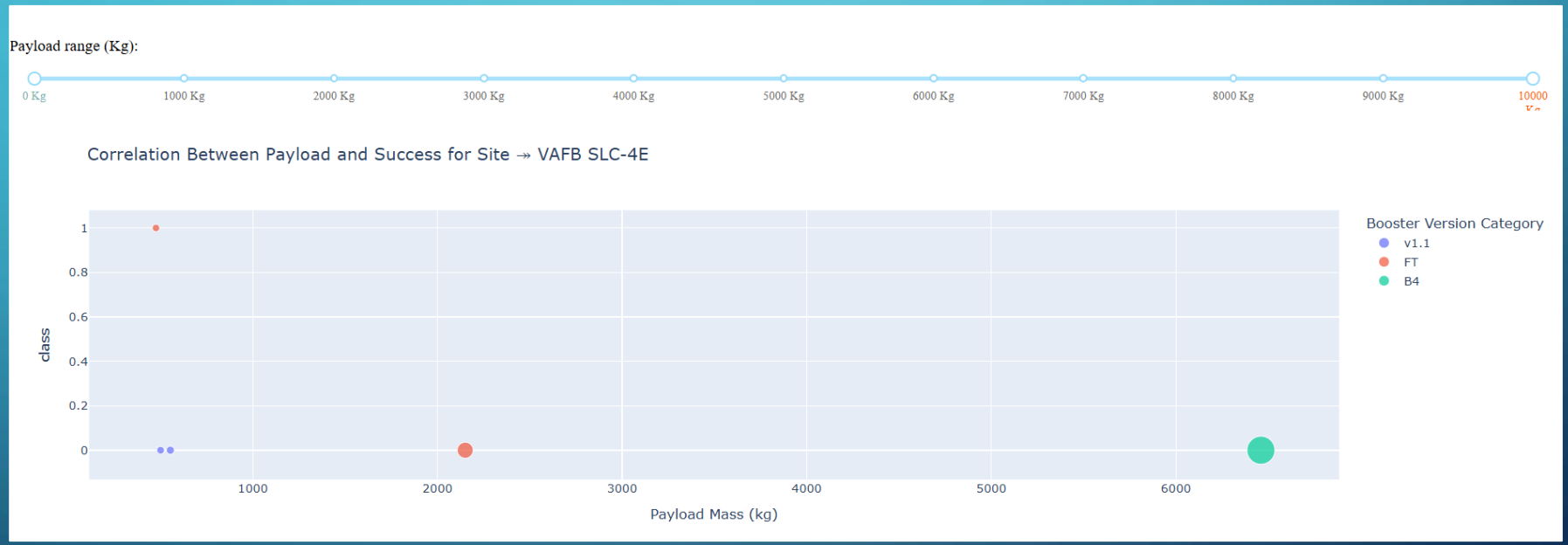
RESULTS – Dash

- Below picture shows a pie chart of launch site VAFB SLC-4E when selected from the dropdown menu.
- From this we can infer that 60% of launches form this site are failed and only 40% are successful (0 – Fail , 1 – Success).



RESULTS – Dash

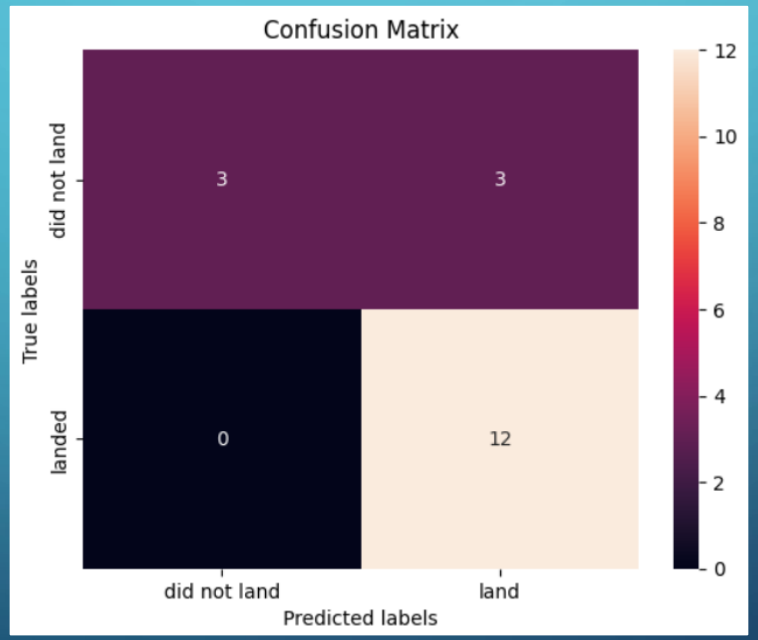
- Below picture shows a scatterplot when payload mass range is set from 0kg to 10000kg.
- Class 0 represents Failed and Class 1represents successful launches.



RESULTS – Predictive Analysis

Logistic Regression:

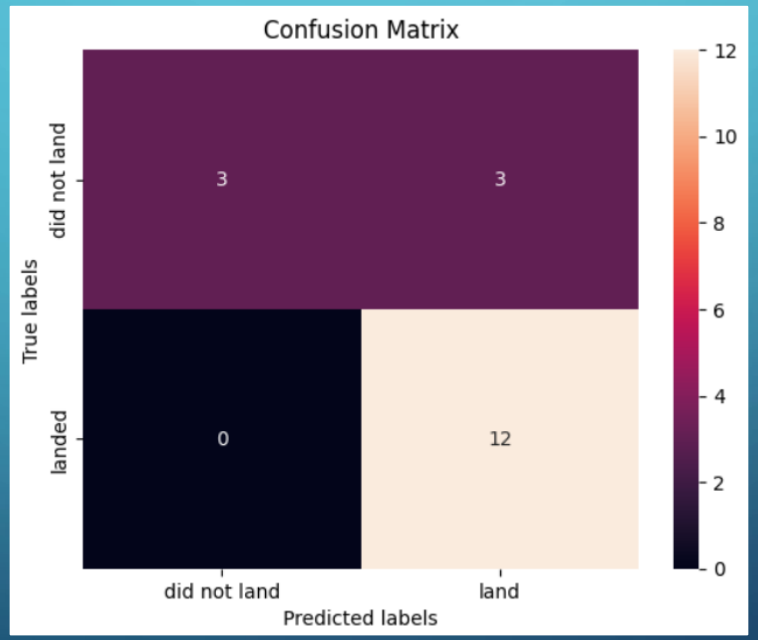
- GridSearchCV best score: 0.8464285714285713
- Accuracy score: 0.8333333333333334
- Confusion matrix:



RESULTS – Predictive Analysis

Support Vector Machine (SVM):

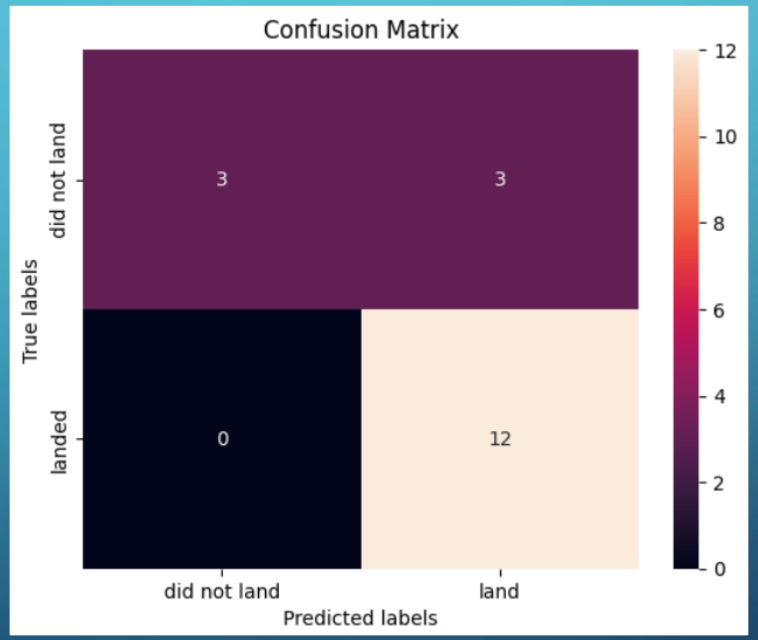
- GridSearchCV best score: 0.8482142857142856
- Accuracy score: 0.8333333333333334
- Confusion matrix:



RESULTS – Predictive Analysis

Decision Tree:

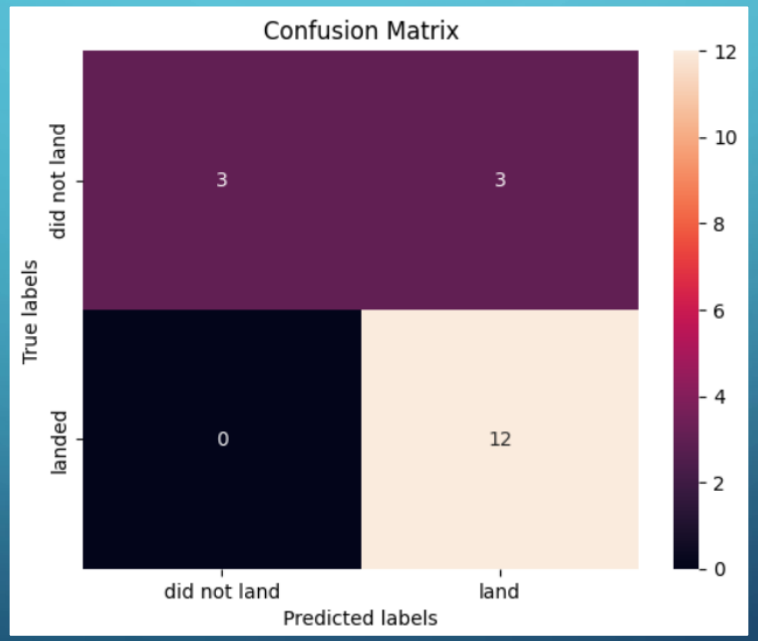
- GridSearchCV best score: 0.8892857142857142
- Accuracy score: 0.8333333333333334
- Confusion matrix:



RESULTS – Predictive Analysis

K Nearest Neighbours (KNN):

- GridSearchCV best score: 0.8482142857142858
- Accuracy score: 0.8333333333333334
- Confusion matrix:



RESULTS – Predictive Analysis

- All the 4 models that we have created and tested share the same accuracy score and confusion matrix.
- Hence, based on GridSearchCV best scores we rank the models in the following order from best to worst.

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

CONCLUSION

- We have successfully determined the best model to predict whether the first stage of SpaceX Falcon 9 launch will land successfully or fail to land in order to determine the cost of launch.
- Low weighted payloads perform much better than heavier payloads.
- The success rates for SpaceX launches is directly proportional to time in years they will eventually perfect the launches.
- KSC LC 39A had the most successful launches from all the sites.
- The predictive model produced by decision tree algorithm performed the best among all the four machine learning algorithms employed.

Thank you!