Winning Space Race with Data Science

Chirag Jain

21.12.2024

Overview

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

Executive Summary

The data was collected with the help of SpaceX API and Wikipedia, which underwent data preprocessing, selection and integration to obtain the required attributes. To understand the data, data visualization libraries and SQL were used for Exploratory Data Analysis. Folium library was used to map the locations of each launch site along with its corresponding successful and unsuccessful launches, and distance from coastline, nearest city and highway. All the visualizations were displayed via a Dash App. Afterwards, comparative predictive models were built to identify the best model to predict a successful launch of a Falcon 9 rocket.

K Nearest Neighbour was identified as the best model to predict the success of a launch from sites with an accuracy of 90.35% on test data. Further improvement in processing algorithm could benefit the model to perform relatively better than now.

Introduction

SpaceX is a billion-dollar private space corporation, specializing in rocket manufacturing and launching. Falcon 9, a commercial rocket known for its reusability, is a medium-lift launch vehicle. In this capstone, we will be conducting a time-series analysis and create a prediction model on the success of its launch based on the dataset acquired via SpaceX API and Wikipedia.

Falcon 9 costs about 62 million dollars, relatively lower than its competitors (165 million dollars) for commercial use. Hence, determining the successful landing rate of the first stage will help give insight to alternative start-ups/MNCs to bid against SpaceX for rocket launches.

The main question, we would be answering is, whether the Falcon 9 rocket lands successfully based on the payload mass, orbit type, launch site and so on ?

Methodology

- Data Collection
  - We acquired the data on SpaceX rockets from the SpaceX API (https://api.spacexdata.com/v4/rockets/), which was then filtered to extract only Falcon 9 rockets.
  - We addressed the missing payload mass values by taking the mean of that attribute and replacing missing values with the mean.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

Fig.1 Acquired Data from SpaceX API

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Fig.2 Dataset after Data Wrangling

Methodology

- Data Scraping
  - We acquired Falcon 9 launch details from this Wikipedia Page . We gathered data on the flight date, launch site, payload details, payload mass, expected orbit, customer, launch outcome, and booster version.
  - We used Request library to access the Wiki page and Beautiful Soup to scrape and parse it, and stored it in a dataframe for further use.

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	[4 June 2010,, 18:45]	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	[4 June 2010,, 18:45]	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0004.18	Failure	4 June 2010	18:45
2	[4 June 2010,, 18:45]	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	4 June 2010	18:45
3	[4 June 2010,, 18:45]	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.07B0006.18	No attempt	8 December 2010	15:43
4	[8 December 2010,, 15:43]	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.07B0007.18	No attempt\n	22 May 2012	07:44

Fig.3 Scraped dataset from Wikipedia

Methodology

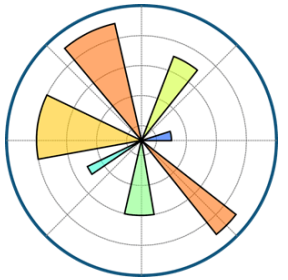
- Data Wrangling
  - In this process, from the launch outcome and booster landing, we are inferring whether the launch was successful or not. On that basis, we are identifying the success rate for each launch site.
  - We create a class consisting of 0's and 1's, indicating whether the launch was successful or not. This will be useful while building the prediction model.
  - In the end, we have a dataset containing 90 rows and 18 attributes, ready for exploratory data analysis.

Methodology

- Exploratory Data Analysis
  - SQL (sqlalchemy) - The data is acquired using SQL query to perform the following task:
    - Identify the names of launch sites.
    - Display records where launch sites' name starts with "CCA".
    - Display total payload mass carried by boosters launched by NASA(CRS)
    - Display average payload mass carried by F9 v1.1 booster
    - List the date when the first successful landing outcome on launch pad had occurred.
    - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - List the total number of successful and failure mission outcomes
    - List the names of the booster versions which have carried the maximum payload mass.
    - List the records which will display the month names, failure landing outcomes in drone ship ,booster versions, launch site for the months in year 2015.
    - 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.

Methodology

- Exploratory Data Analysis
  - Matplotlib and Seaborn – Tasks:
    - Visualize the relationship between Flight Number and Launch Site
    - Visualize the relationship between Payload Mass and Launch Site
    - Visualize the relationship between success rate of each orbit type
    - Visualize the relationship between Flight Number and Orbit type
    - Visualize the relationship between Payload Mass and Orbit type
    - Visualize the launch success yearly trend



- Feature Engineering
  - In this, we discarded irrelevant attributes and reformatted certain attributes to 0's and 1's via one hot encoding.

	FlightNumber	PayloadMass	Orbit	LaunchSite	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
0	1	6104.959412	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0003
1	2	525.000000	LEO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0005
2	3	677.000000	ISS	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B0007
3	4	500.000000	PO	VAFB SLC 4E	1	False	False	False	NaN	1.0	0	B1003
4	5	3170.000000	GTO	CCAFS SLC 40	1	False	False	False	NaN	1.0	0	B1004

Fig.4 Dataset with selected features before Once Hot Encoding

Methodology

- Building an Interactive Map Using Folium and Plotly Dash
  - We used folium library to build an interactive map depicting the locations of launch sites along with markers indicating successful and unsuccessful launches.
  - We also used Dash application to create a frontend to access the visualizations and interact with it.

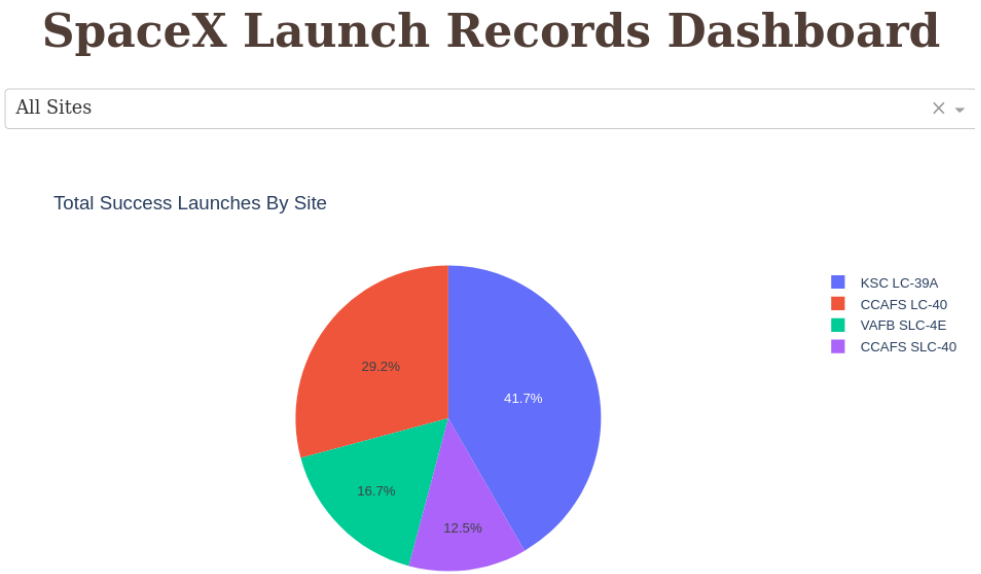


Fig.5 SpaceX launch Records Dashboard built using Dash

Methodology

- Machine Learning Prediction Model
  - We built 4 models and conducted a comparative analysis on them to identify the best out of the 4. The models selected were – Logistic Regression, Section Vector Machine (SVM), Decision Tree, and K Nearest Neighbour (KNN).
  - We standardised the dataset using Standard Scalar function, and separated the dataset into training set and testing set in the ratio of 8:2 with a 2 as the random state value.
  - Each model was subjected to GridSearchCV to identify the best parameters for that model with Cross-validation set for 10-folds.
  - The models were done compared against each other based upon their accuracy. KNN model came on top with the best test score (90.35%), where as LR and SVM topped on the training set (83.33%).

Results

EDA using SQL

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

Task 1: Unique Launch Sites

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

Task 2: Displaying 5 records where launch sites' begin with "CCA"

Total_Payload_Mass
45596

Task 3: Displaying total payload mass carried by boosters launched by NASA (CRS)

Average_F9_Payload_Mass
2534.6666666666665

Task 4: Displaying average payload mass carried by boosters (F9 v1.1)

min(Date)
2015-12-22

Task 5: List the date when the first successful landing outcome on launch pad had occurred.

Results

EDA using SQL

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

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

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

Task 9: Listing the records with Month, Landing Outcome in Drone Ship, Booster Version, Launch Site in the year 2015

Successful	Failure
100	1

Task 7: Listing total number of successful and failure mission outcomes

Landing_Outcome	Outcome_Count
No attempt	21
Success (drone ship)	14
Success (ground pad)	9
Failure (drone ship)	5
Controlled (ocean)	5
Uncontrolled (ocean)	2
Precluded (drone ship)	1

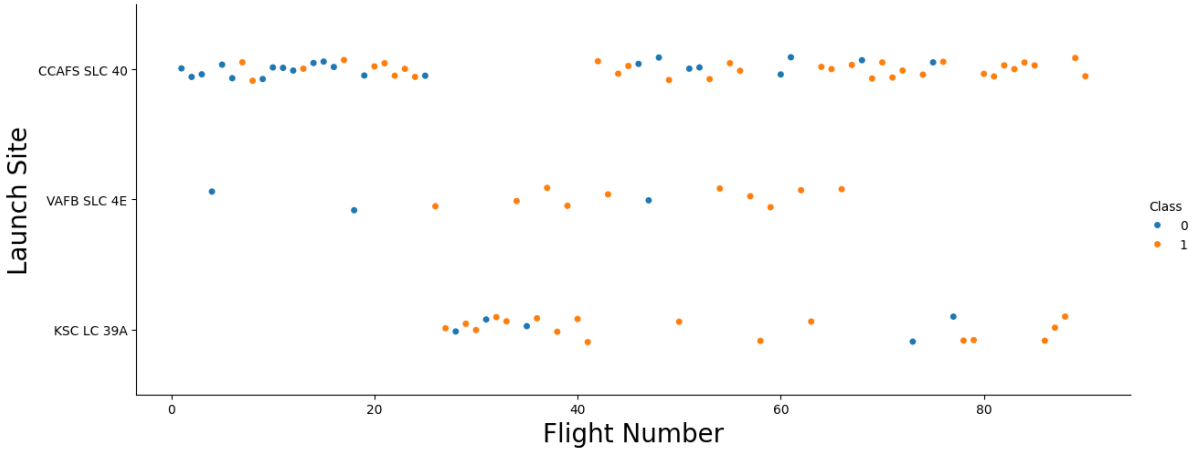
Task 10: Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

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

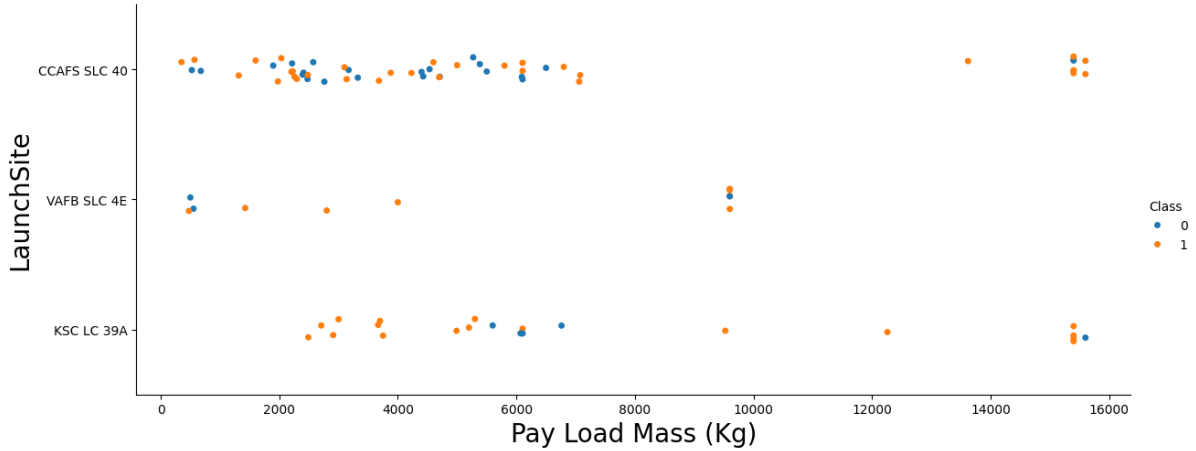
Task 8: Listing the boosters carrying the maximum payload mass.

Results

EDA using Matplotlib and Seaborn



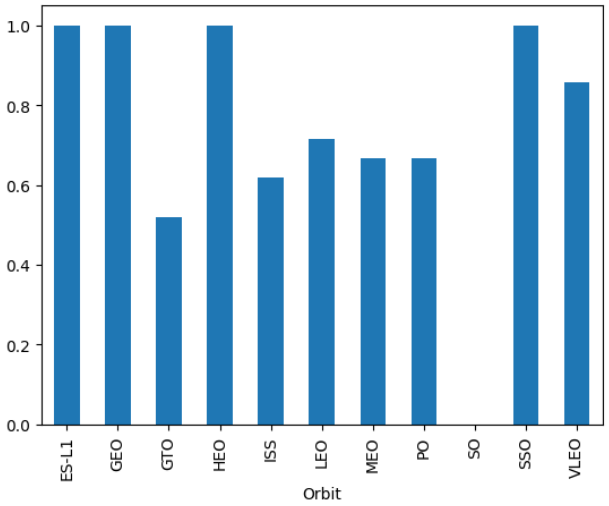
Task 1: Relationship between Flight Number and Launch Site



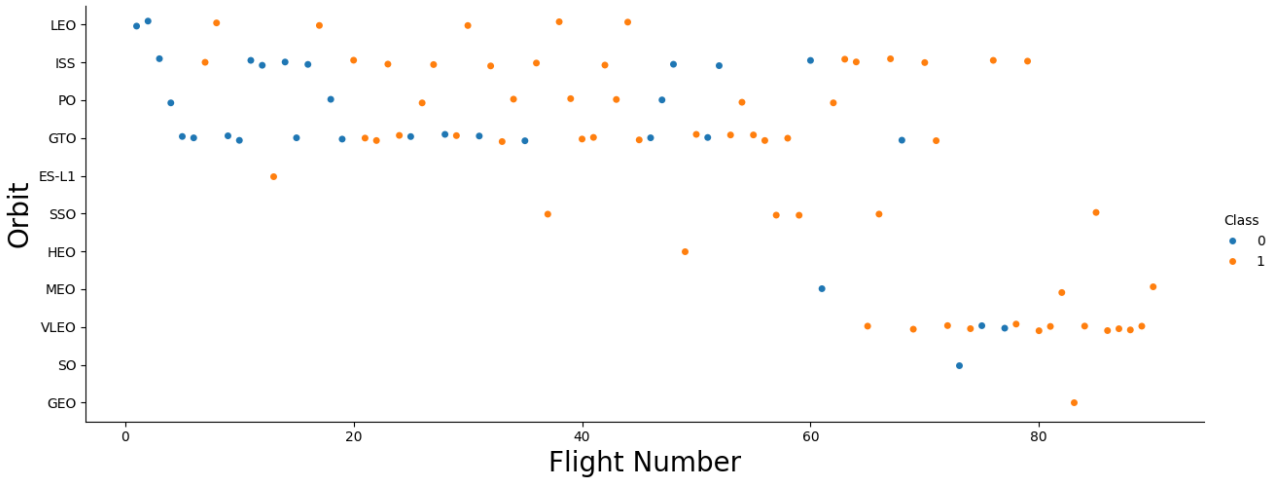
Task 2: Relationship between Payload Mass and Launch Site

Results

EDA using Matplotlib and Seaborn



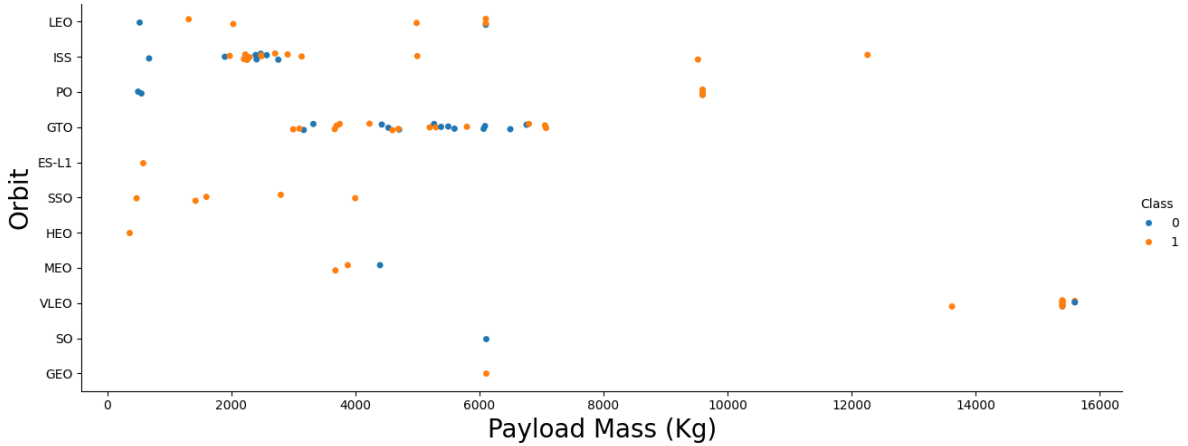
Task 3: Visualising success rate of each orbit type



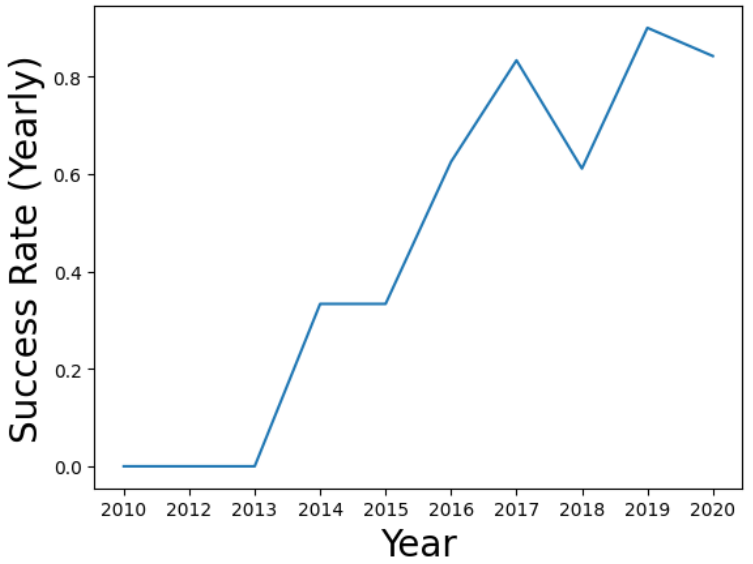
Task 4: Relationship between Flight Number and Orbit type

Results

EDA using Matplotlib and Seaborn



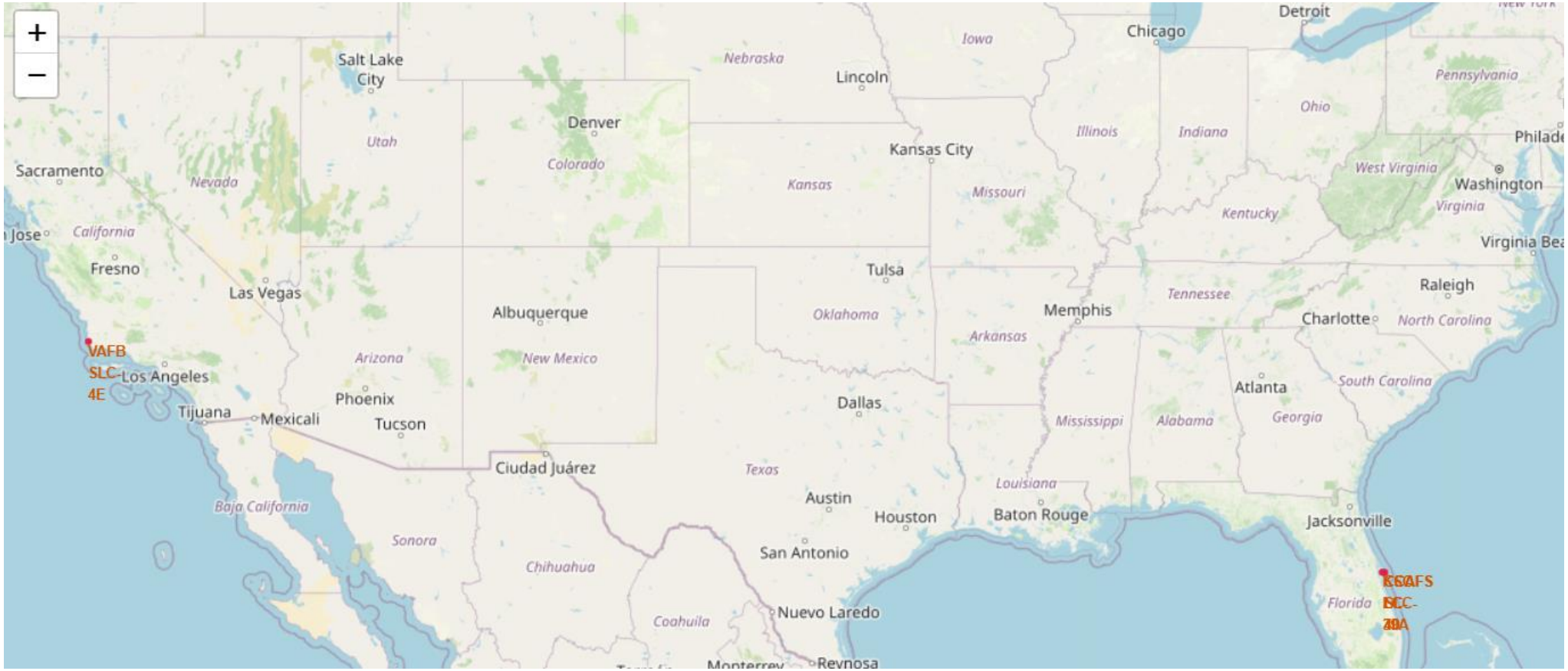
Task 5: Relationship between Payload Mass and Orbit Type



Task 6: Visualising launch success yearly trend

Results

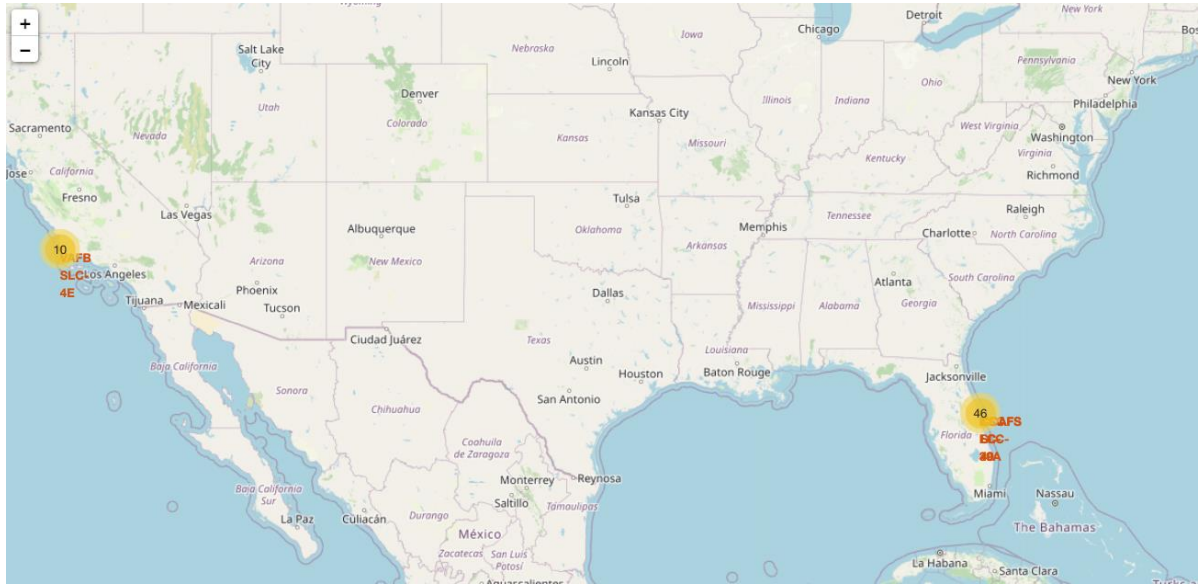
Interactive Visual Analytics using Folium

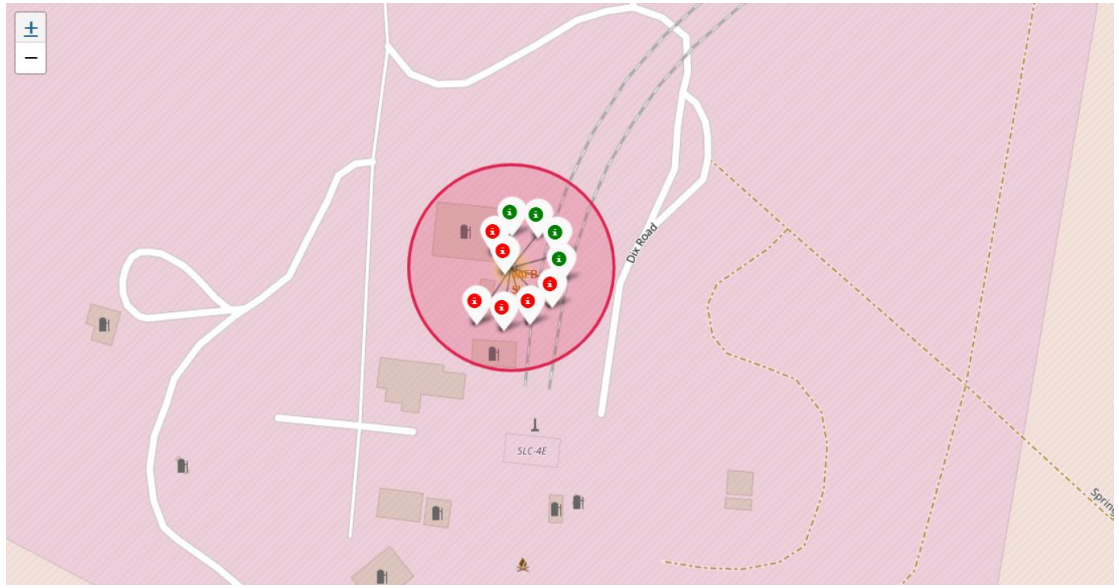


Task 1: All Launch Sites on the Map

Results

Interactive Visual Analytics using Folium

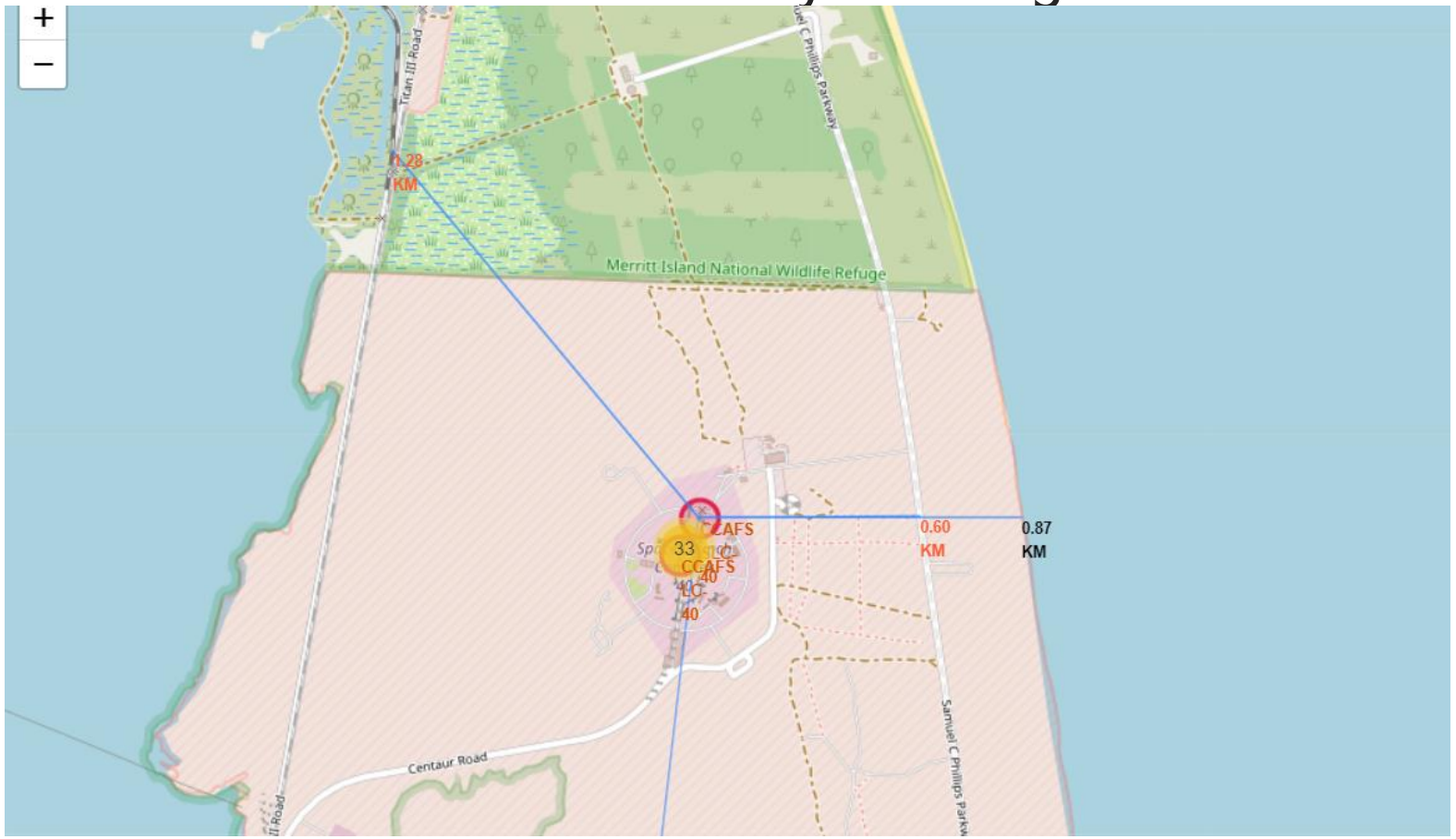




Task 2: Marking Successful and Failed Launches for all launch sites

Results

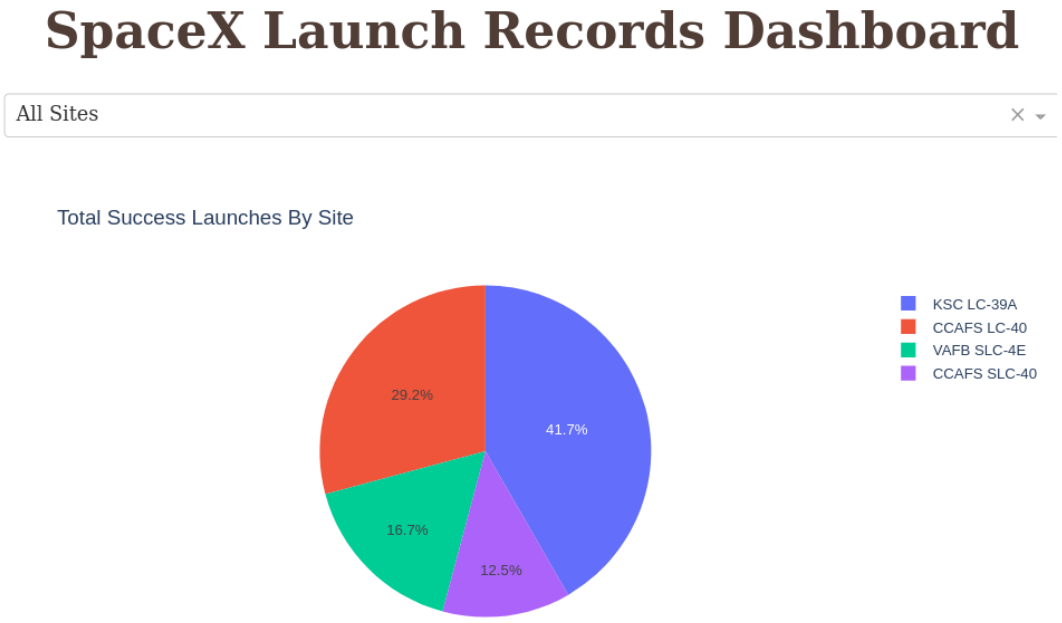
Interactive Visual Analytics using Folium



Task 3: Calculating Distance between launch sites and its proximities with Polylines

Results

Dash

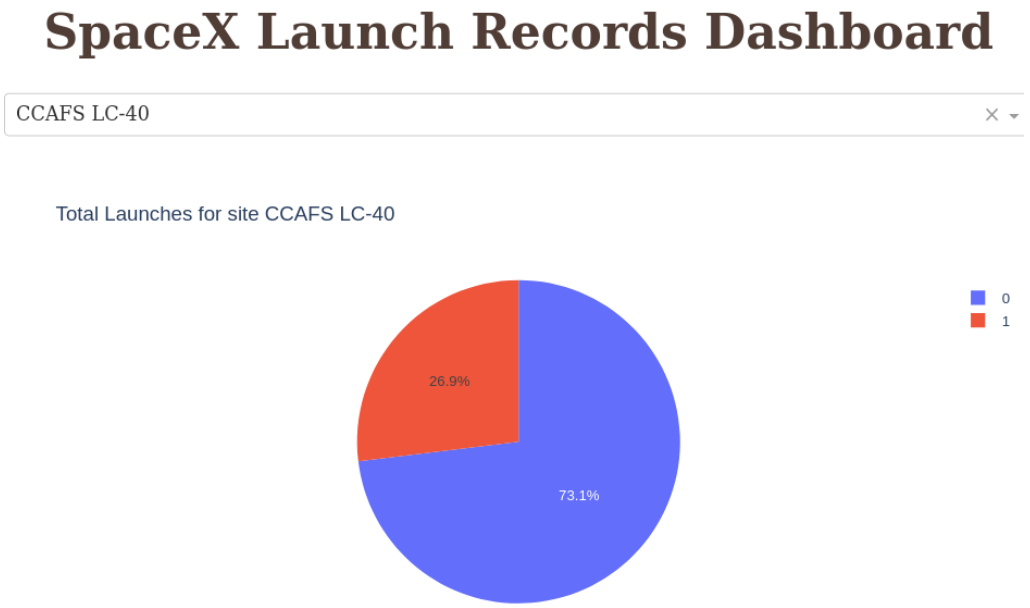


Displaying Total Launches by Each Site via a Pie Chart

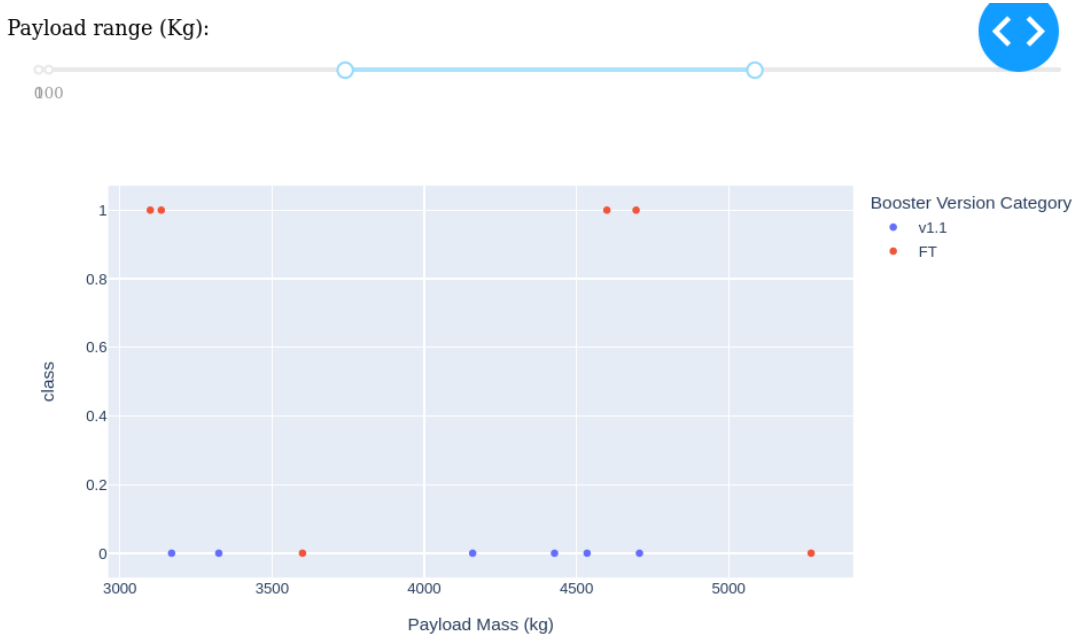
Interactive visualisation with varying Payload Range

Results

Dash



Displaying Total Launches by CCAFS LC-40 Site

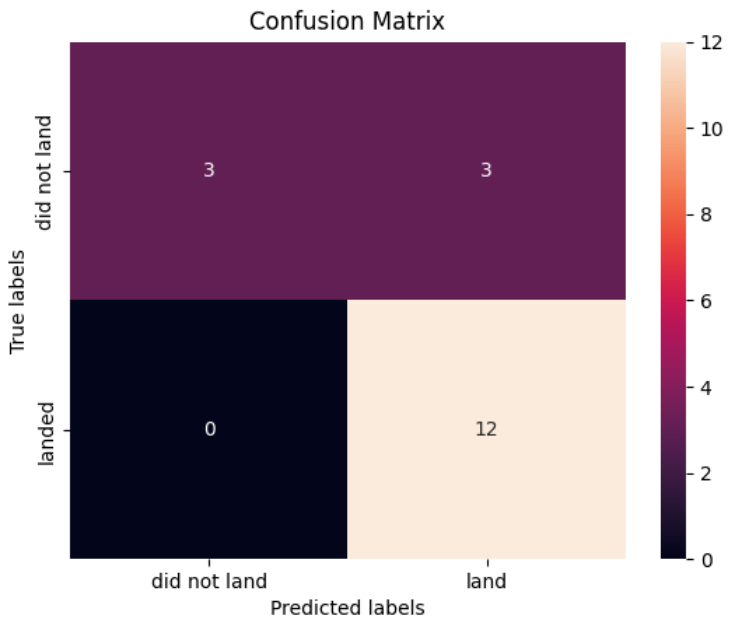


Displaying booster classes carrying payload between 3000Kgs and 5500 Kgs

Results

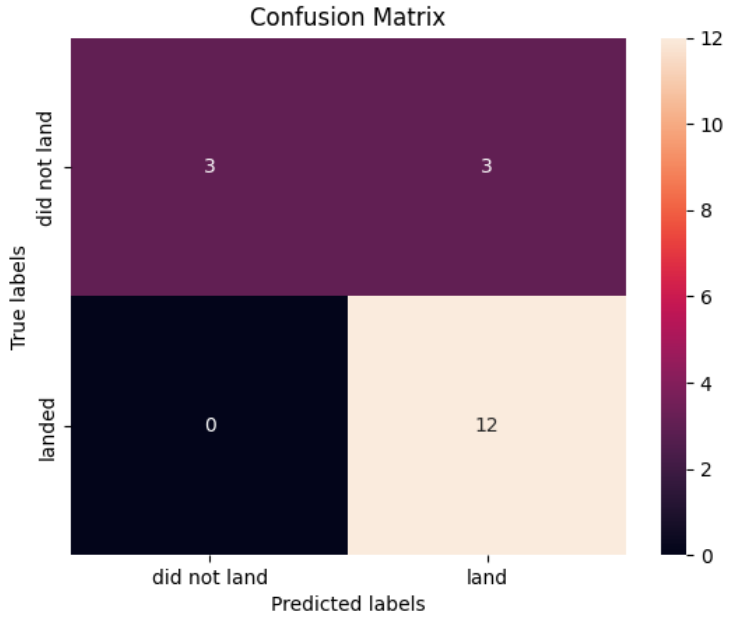
Predictive Model

Logistic Regression:
Training Accuracy : 83.33%
Test Accuracy : 84.64%



Logistic Regression Confusion Matrix

Section Vector Machine(SVM):
Training Accuracy : 83.33%
Test Accuracy : 84.82%

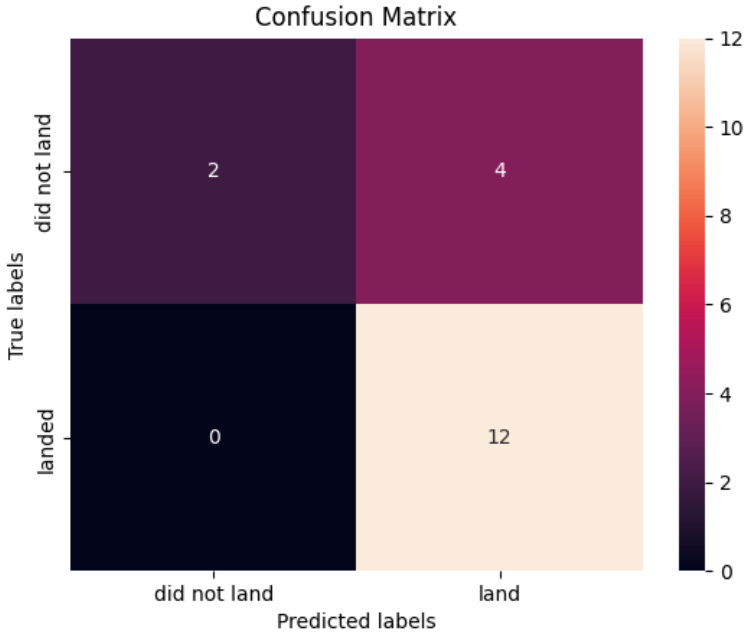


SVM Confusion Matrix

Results

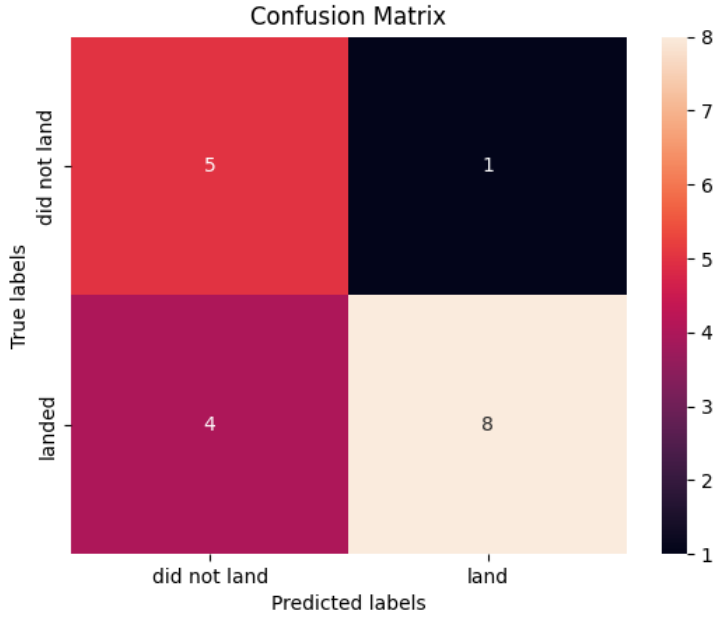
Predictive Model

Decision Tree:
Training Accuracy : 77.77%
Test Accuracy : 88.92%



Decision Tree Confusion Matrix

K Nearest Neighbour (KNN):
Training Accuracy : 72.22%
Test Accuracy : 90.35%



KNN Confusion Matrix

Discussion

From the results it could be observed that there might be some relation among the features with the mission outcome. For example, with heavy payloads the successful landing or positive landing rate are more for orbit types Polar, LEO and ISS. However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

With further improvement in the algorithm used for pre-processing techniques and to train the machine learning models, we could obtain better insight into correlation among the features. Techniques like correlation analysis, data bivariate and multivariate analysis, and data augmentation may prove to improve our understanding of the data and help in knowledge discovery.

Conclusion

- In this Capstone, we tried to identify the mission outcomes of the Falcon9 SpaceX rockets based on environmental and rocket attributes.
- From the analysis of data it was identified that site KSC LC-39A had the best record of successful launches and that launches with Payload mass above 7000Kgs were less risky.
- Furthermore, an increasing trend of successful launches can be observed from 2013 onwards, with a slight dip in 2018.
- From the predictive model, KNN proved to be the best to fit on the test data to identify potential of success for a launch mission.