SpaceX Falcon 9 first stage Landing Prediction

Data Science Capstone project

ASHFAQ H

(SEPTEMBER 2021)

Outline

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

Executive Summary

The goal of this final capstone project is to develop a data product using the machine learning algorithms that will predict the successful landing of the first stage of the SpaceX's Falcon 9 rocket. The data published in the SpaceX website is used to for this project which is extracted using API requests sent to SpaceX website.

This is the project report to demonstrate the completion of a milestone of the capstone final project. In this report, the steps taken to acquire, load, data cleaning, exploratory analysis, visual analysis and predictive analysis done on the data are documented.

Introduction

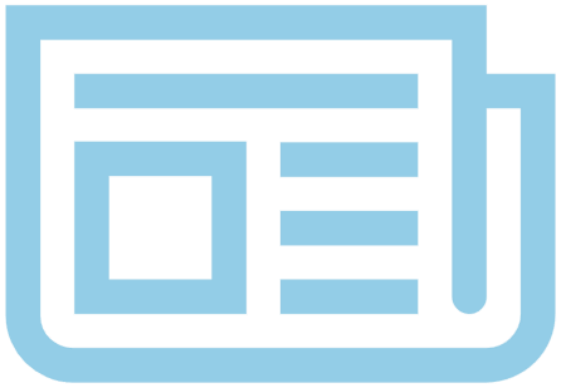
Project background and context

SpaceX advertises Falcon 9 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.

Problem Statement

If it can be determined if the first stage will land, cost of the launch can be determined. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Methodology



- Data collection methodology:
  - Data was collected from SpaceX websites using SpaceX APIs
  - Booster version, Launch site, Payload data and Core data were obtained from SpaceX APIs
- Perform data wrangling
  - Normalizing the Json data, keeping only the required columns and required booster version i.e., Falcon9.
  - Null value handling, replacing the missing values with suitable values
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic regression, KNN, Decision trees and SVM algorithms were used for training and prediction

Methodology

Since the success of landing of Falcon 9 rocket was to be predicted, various features contribute for the success or failure of the Rocket landing. Hence the data pertaining to Booster version, Launch site, Payload data and Core data were extracted from SpaceX APIs. After the required data was collected, EDA was performed using visualization and SQL. The correlation between the features were analyzed by interactive visual analytics using Folium and Plotly Dash.

Post the data analysis, Logistic regression, KNN, Decision trees and SVM algorithms were used for training and prediction of the extracted data and based on which the conclusion of the data is provided.

Tools Used: Web-scraping of SpaceX site was done to consolidate data-frame information which was saved as csv files for convenience and to simplify the report. Geodata was obtained by coding a program to use Nominatim to get latitude and longitude of subway stations and for each of (144 units) the apartments for rent listed. Geopy_distance and Nominatim were used to establish relative distances. Seaborn graphic was used for general statistics on rental data. Maps with popups labels allow quick identification of location, price and feature, thus making the selection very easy

Data collection

Describe how data sets were collected.

a. Data was collected by SpaceX API using GET Request and the collected data in Json is converted to Pandas DataFrame
b. After the data is normalized, a separate data frame is created using a set of features
c. The data is filtered for single core, single payloads and date is converted to standard format
d. Booster version is filtered for Falcon 9 rockets
e. Missing values are handled by replacing missing values by mean values for payload mass

Data collection – SpaceX API

The SpaceX requested API is used to extract information using identification numbers in the launch data

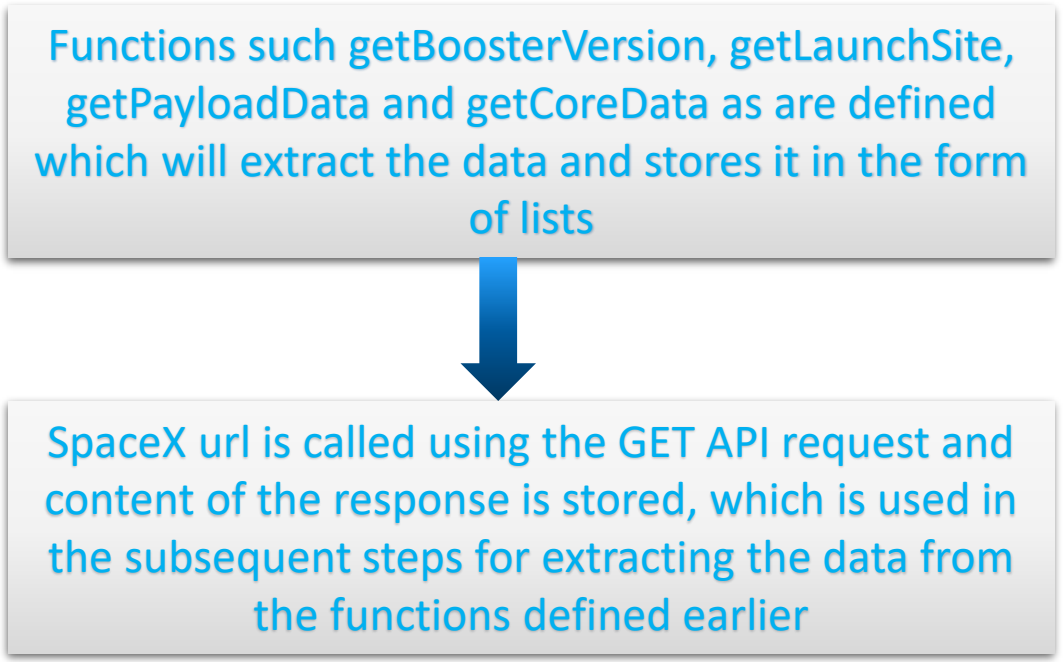
The functions that were used to extract the data from json are getBoosterVersion, getLaunchSite, getPayloadData, getCoreData

The GitHub URL of the completed SpaceX API calls notebook -

Data collection notebook

(Click on the hyperlink to follow)

Flowchart of SpaceX API calls here



Data collection – Web scraping

a. From the **rocket** we would like to learn the booster name
b. From the **payload** we would like to learn the mass of the payload and the orbit that it is going to
c. From the **launchpad** we would like to know the name of the launch site being used, the longitude, and the latitude.
d. From **cores** we would like to learn the outcome of the landing, the type of the landing, number of flights with that core, whether gridfins were used, whether the core is reused, whether legs were used, the landing pad used, the block of the core which is a number used to separate version of cores, the number of times this specific core has been reused, and the serial of the core.

The GitHub URL of the completed web scraping notebook, as an external reference - EDA Notebook (Click on the hyperlink to follow)

Data wrangling

- Describe how data were processed
  a. The Columns in the data where the missing values were identified
  b. The LandingPad column containing the Null values were retained since they represented where the landing pads were not used
  c. Mean was calculated for the PayloadMass column which had Null values and the mean value was replaced with Null values
- You need to present your data wrangling process using key phrases and flowcharts

The GitHub URL of the completed data wrangling related notebooks, as an external reference - EDA with SQL Notebook (Click on the hyperlink to follow)

EDA with data visualization

- Summarize what charts were plotted and why used those charts
    a. Multiple scatter plots, plot and line plot were plotted again the attributes to visualize the launch outcome.
    b. Scatter charts such as FlightNumber vs. PayloadMass, FlightNumber vs. LaunchSite, launch sites vs. their payload mass, FlightNumber and Orbit type and Payload vs. Orbit were plotted, and the relationship were established
    c. Bar plot was plotted for check the relation between orbit vs. class
    d. A line chart with x axis to be Year and y axis to be average success rate, was plotted to get the average launch success trend over the years

The GitHub URL of the completed EDA with data visualization notebook, as an external reference - EDA with Data Visualization (Click on the hyperlink to follow)

EDA with SQL

- Summarize performed SQL queries using bullet points
  a. Names of the unique launch sites in the space mission
  b. 5 records where launch sites begin with the string 'CCA’
  c. The total payload mass carried by boosters launched by NASA (CRS)
  d. Average payload mass carried by booster version F9 v1.1
  e. The date when the first successful landing outcome in ground pad was achieved
  f. The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  g. The total number of successful and failure mission outcomes
  h. The names of the booster_versions which have carried the maximum payload mass
  i. The failed landing_outcomes in drone ship, their booster versions, and launch site names for the in year 2015
  j. 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 were ranked

The GitHub URL of the completed EDA with SQL notebook, as an external reference - EDA with SQL (Click on the hyperlink to follow)

Building an interactive map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  a. folium.Circle and folium.Marker methods were used to add a highlighted circle area with a text label on a specific coordinate for each launch site on the site map
  b. A new column in launch_sites dataframe called marker_color was used to store the marker colors based on the class value
  c. The distance between two points on the map can be calculated based on their Latitude and Longitude values
- The Objects were added to analyze:
  - Marker clusters can be a good way to simplify a map containing many markers having the same coordinate
  - MousePosition was added on the map to get coordinate for a mouse over a point on the map so that coordinates of points of interests can be obtained and line can be drawn so that distance can be calculated

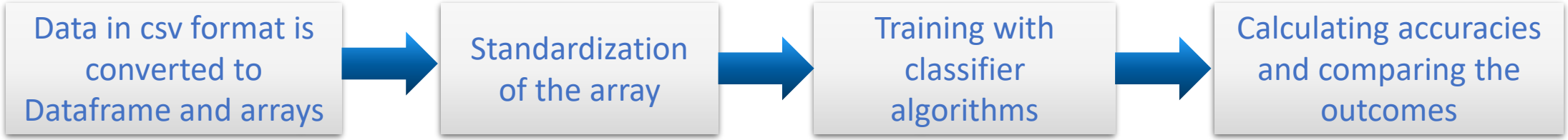
The GitHub URL of the completed interactive map with Folium map, as an external reference - Visual Analytics Notebook (Click on the hyperlink to follow)

Building a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  a. A red circle at NASA Johnson Space Center's coordinate with an icon showing its name
  b. Each of the launch sites were added to the map and displayed
  c. Launch sites with ‘Success’ outcome were indicated in Green color and the launch sites with ‘Failure’ outcome were indicated in Red color in the map and displayed as clusters
- The plots and interactions were added to analyze:
  - If the launch sites are in proximity to equator line
  - If the launch sites are near to coast
  - If the launch sites were close to Highways or railways and their distances

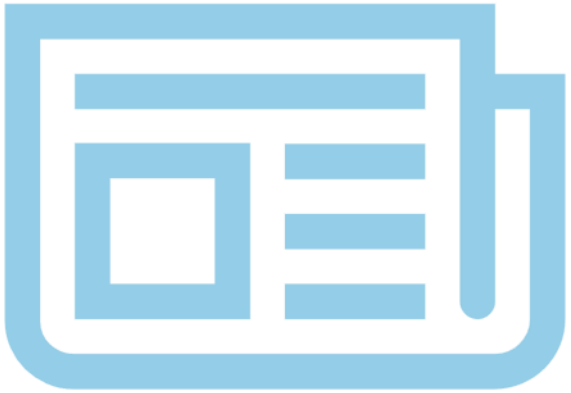
Predictive analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model
  a. The curated data in csv format was loaded into variables to ease the process of performing actions on the data
  b. The standardized data was split to training and test data sets
  c. The data was fed to Logistic regression, SVM classifier, Decision tree Classifier, KNN Algorithms for training on the training data
  d. The corresponding accuracies were computed with test data and the results were compared against each other to determine the best algorithm
- You need present your model development process using key phrases and flowchart



The GitHub URL of the completed predictive analysis lab, as an external reference purpose - Machine Learning Prediction (Click on the hyperlink to follow)

Results



- Exploratory data analysis results:

  Data obtained from the SpaceX API were analyzed and relation between the attributes were recorded wherein the mission outcomes were correlated with orbit type and launch sites columns

- Interactive analytics demo in screenshots:

  Based on the results obtained from the interactive analytics demo, it was observed that most of the stations were near the coastal regions.

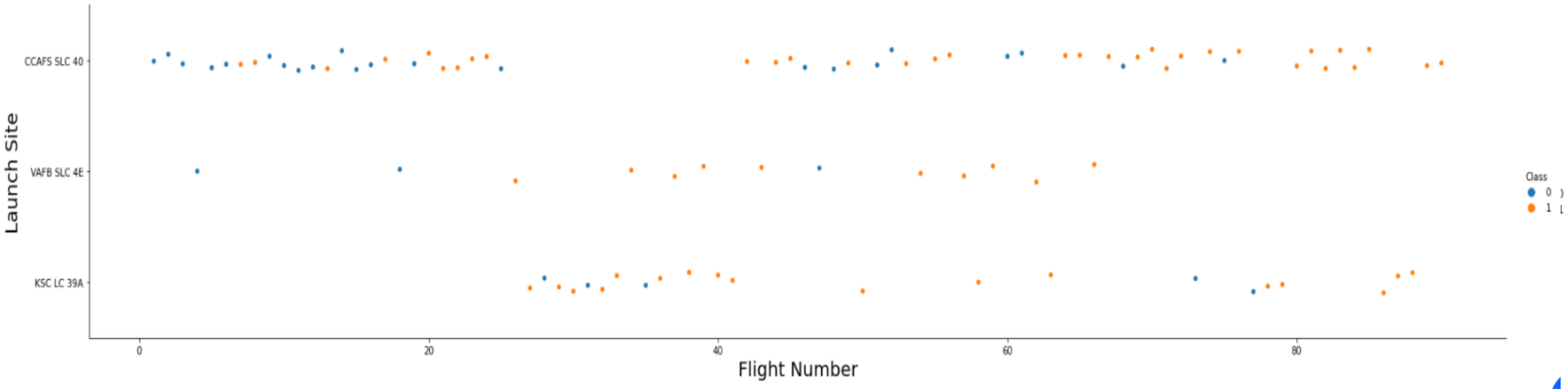
- Predictive analysis results:

  The accuracies were calculated on the test data set, wherein the accuracy of the Decision Tree algorithm is the highest with 88.9% and KNN and Support Vector Machine algorithms having next best accuracies if 84.8%.

EDA with Visualization

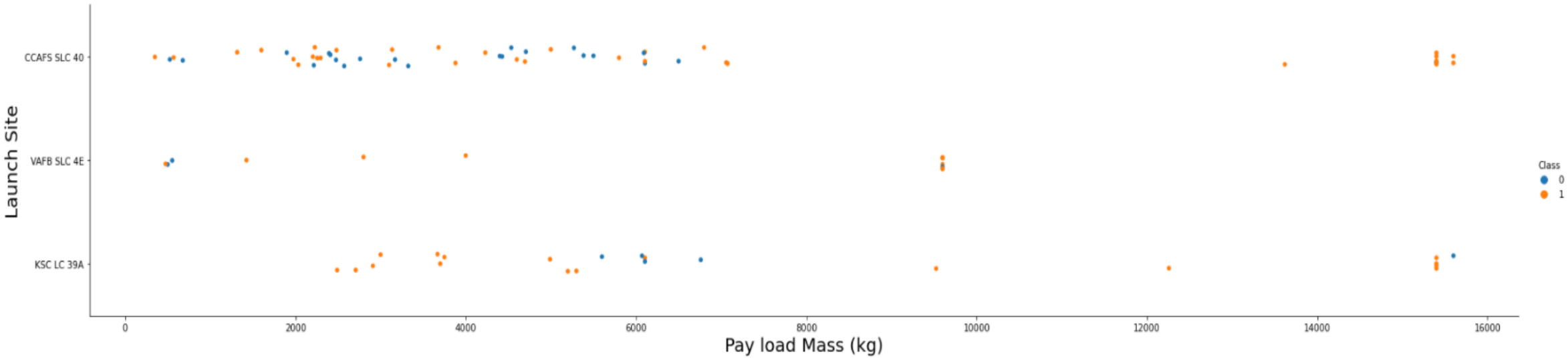
Exploratory Data Analysis was done on the SpaceX data and visualizations were plotted using scatter, line and bar plots to analyze and draw meaningful conclusions on the data

Flight Number vs. Launch Site



We can plot out the FlightNumber vs. Launchsite overlay the outcome of the launch. We see that as the flight number increases, the launch sites with value CCAFS SLC 40 is having higher chance being successful

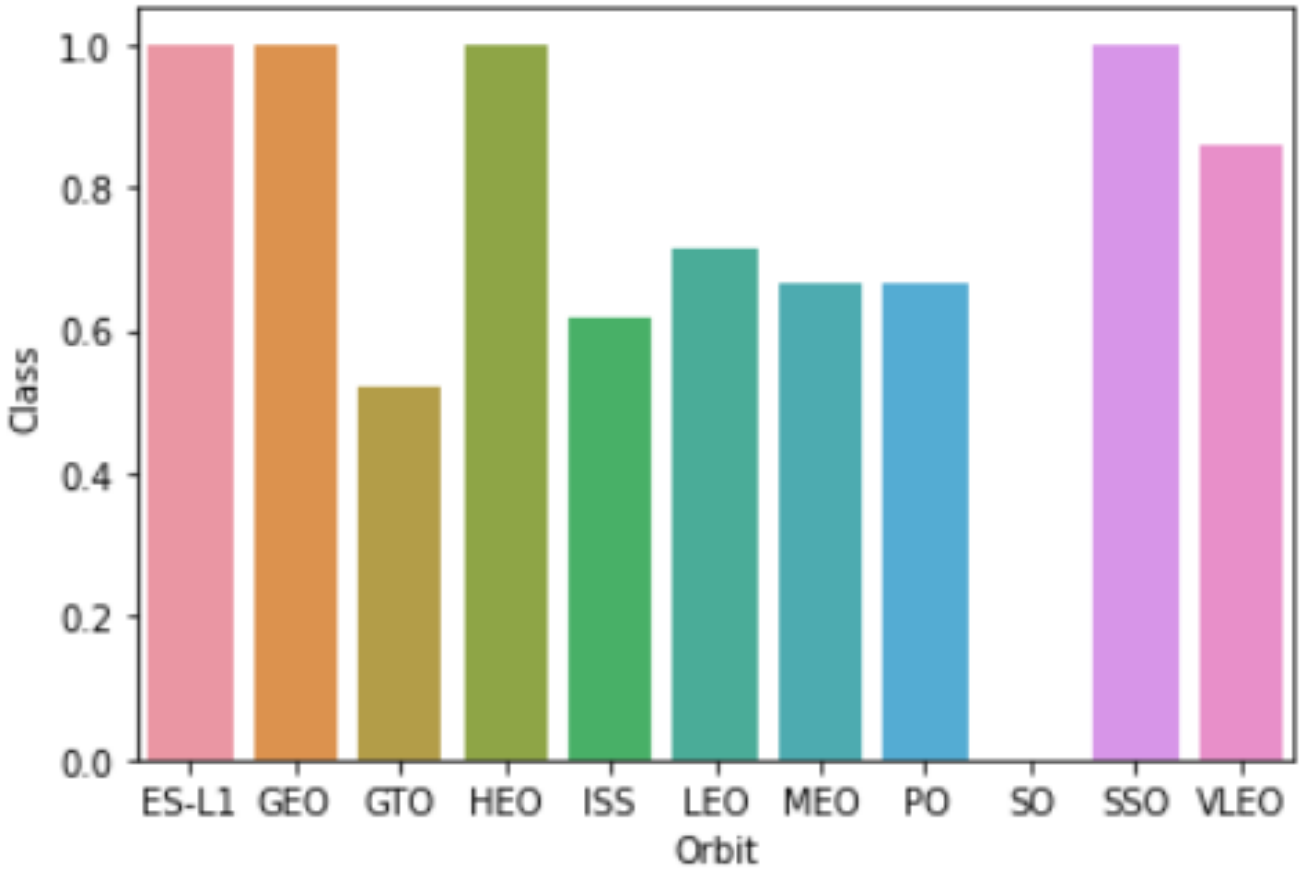
Payload vs. Launch Site



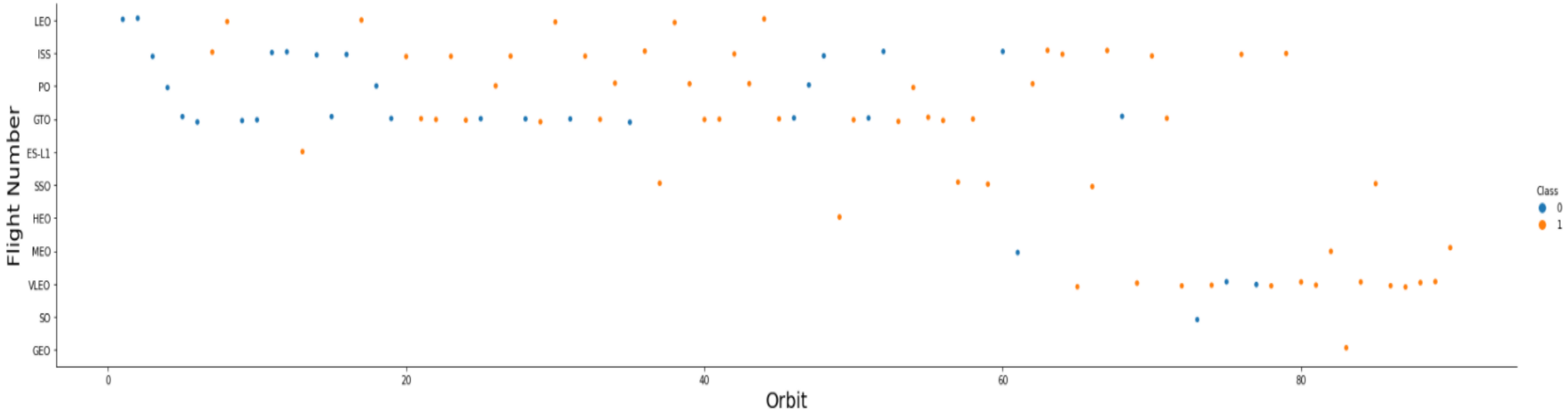
It is observed from the above chart that Launches are most likely to be successful with Pay Load mass above 15000 at sites CCAFS SLC 40 and KSC LC39A.

Success rate vs. Orbit type

It is observed that the orbits such as ES-L1, GEO, HEO and SSO have higher success rate compared to other orbits

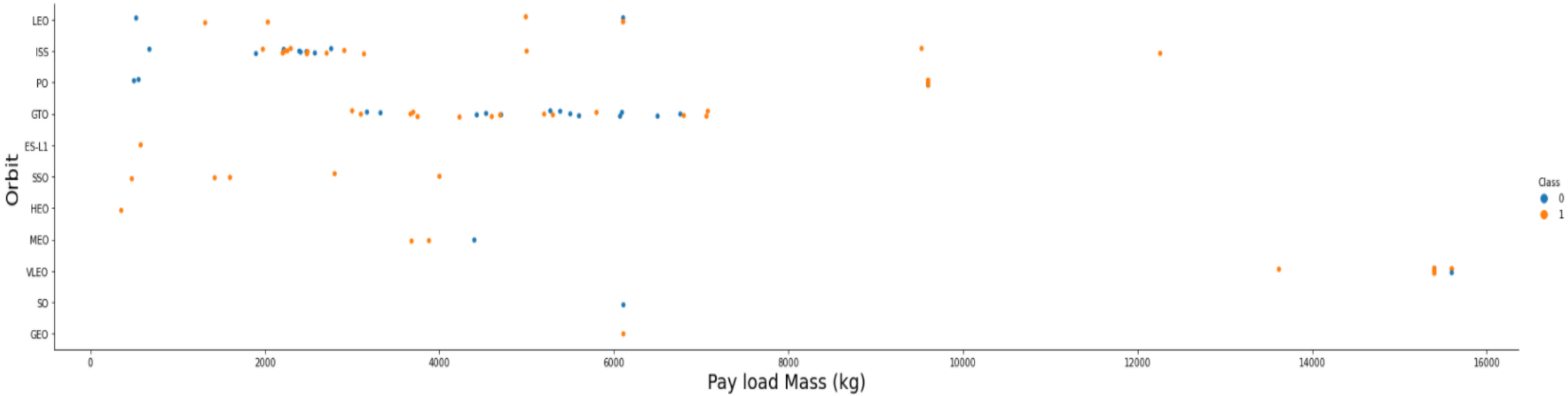


Flight Number vs. Orbit type



It is observed that the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

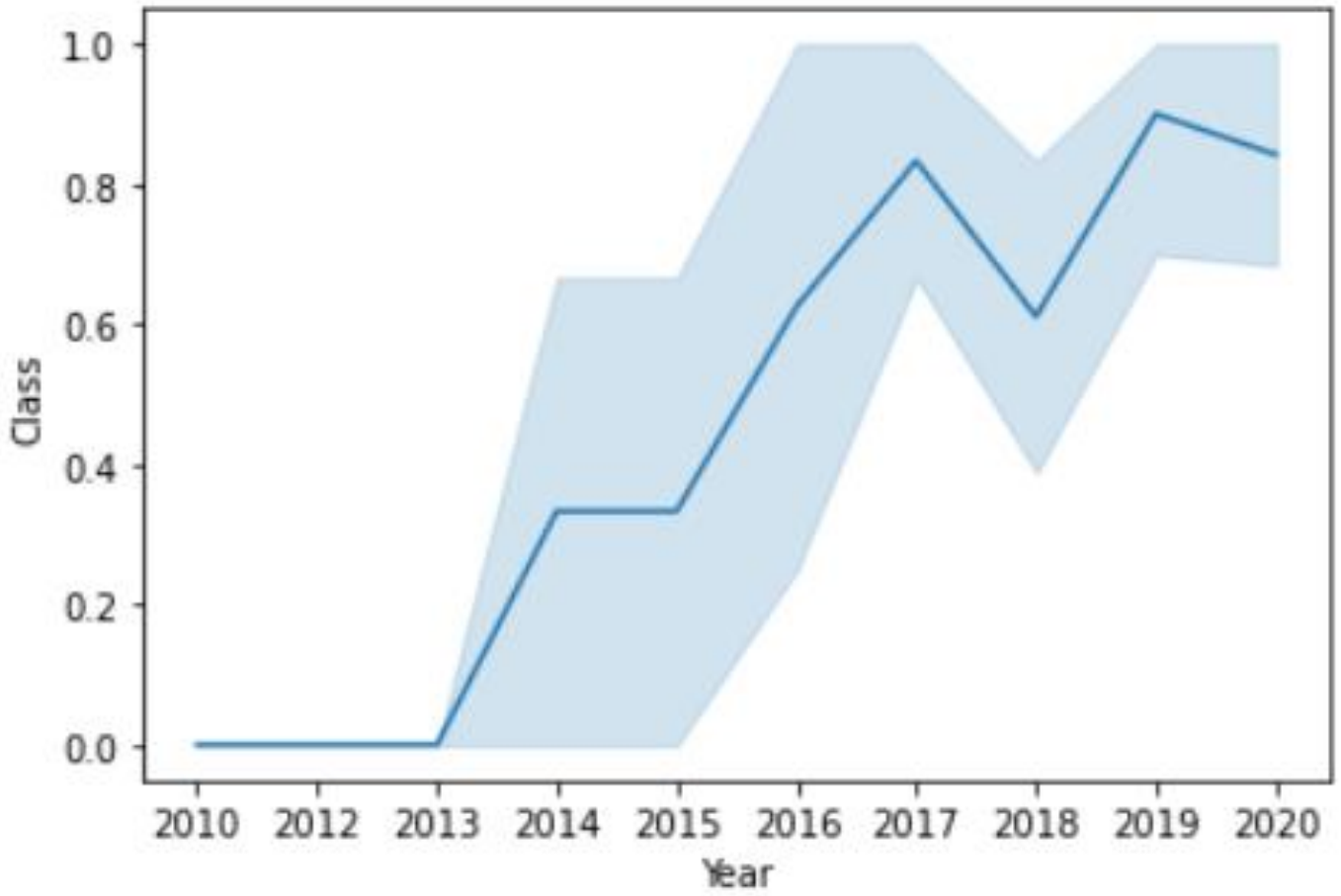
Payload vs. Orbit type



It is observed that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch success yearly trend

It is observed that the success rate since 2013 kept increasing till 2020



EDA with SQL

Exploratory Data Analysis was performed by filtering data to find patters and relationships between the attributes in the data using SQL

All launch site names

The names of the unique launch sites are;

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

Distinct function was run on Launch site column in SpaceX table and 4 launch sites were obtained

Launch site names begin with `CCA`

All launch sites beginning with `CCA`

- CCAFS LC-40
- CCAFS SLC-40

The query used to find the all the launch sites beginning with CCA was

“SELECT * FROM SPACEXTBL WHERE Launch_Site LIKE 'CCA%’ ” for which the 60 rows of data was obtained as the outcome and unique values of the launch sites beginning with CCA is 2.

Total payload mass

The total payload carried by boosters from NASA was 619967 kg

The mass of the payloads range from 0 kg to several rockets carrying over 15000 kg.

Average payload mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928 kg of mass

The query used to find the average payload mass carried by booster F9 v1.1 is as below;

“SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTBL WHERE Booster_Version='F9 v1.1’ “

First successful ground landing date

The date when the first successful landing outcome in ground pad was on 22nd December 2015.

The query used to find the first successful ground landing date is;

“SELECT MIN(Date) FROM SPACEXTBL WHERE Landing_Outcome='Success (ground pad)’ “ for which the min function was used.

Successful drone ship landing with payload between 4000 and 6000

The names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000 are listed below;

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

When the sql query was used to find the result, 4 outomes that were listed was displayed, the query used was

“SELECT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome='Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 “

Total number of successful and failure mission outcomes

The total number of successful and failure mission outcomes are as follows;

- Successful – 100
- Failure - 1

Out of 101 missions, 100 were successful, for one mission payload status was unclear and the rest were successful.

Boosters carried maximum **payload**

The names of the booster which have carried the maximum payload mass were;

- 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

Interactive map with Folium

Interactive maps using folium were plotted to be able to find some geographical patterns about launch sites

All the Launch Sites on the Map

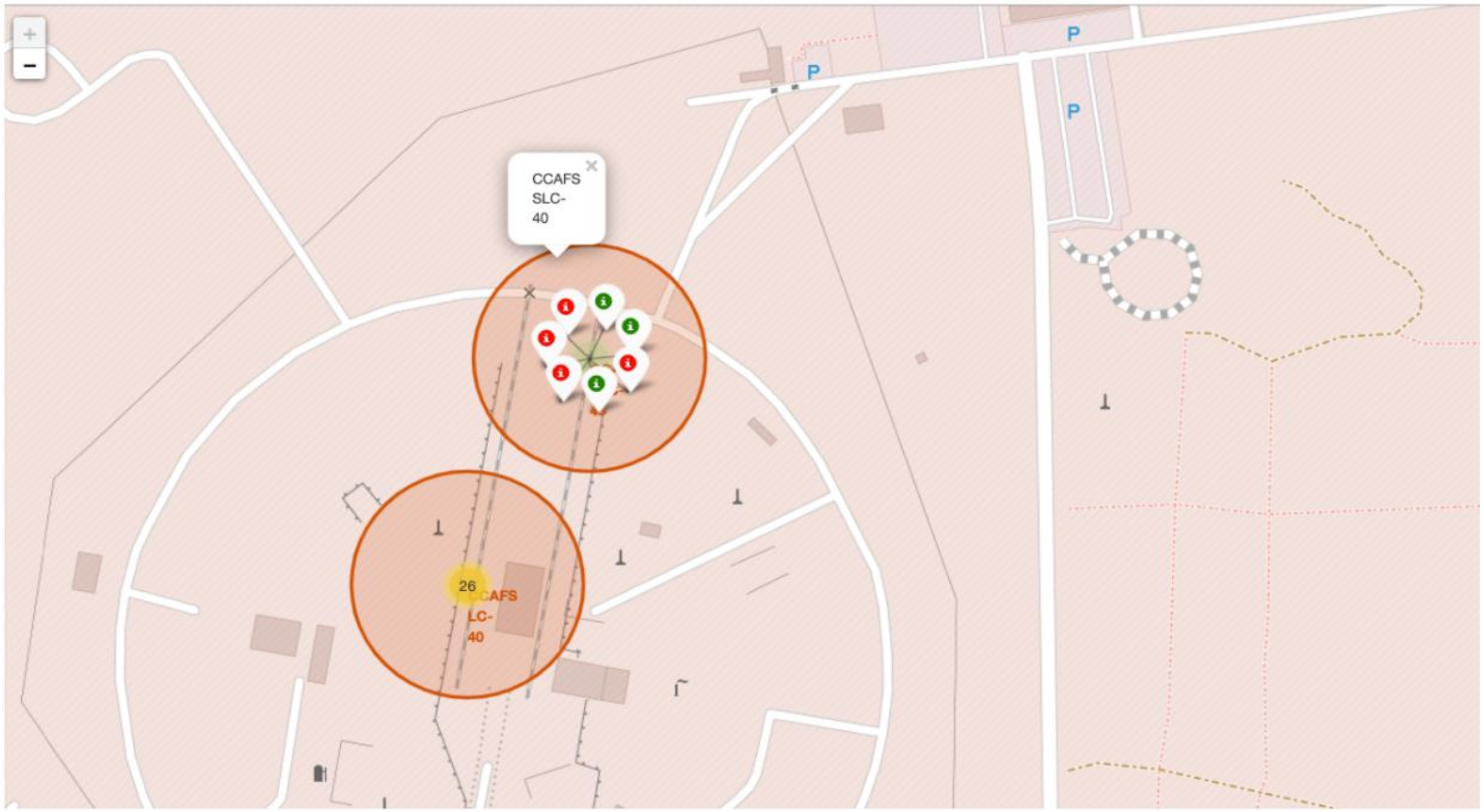
All the launch sites are displayed on a folium map, they are;

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

Success/Failed launches for each site

All the launch sites in the form of clusters are displayed such that green represents ‘success’ and red represents ‘failure’ of the mission.

From the color-labeled markers in marker clusters, one should be able to easily identify which launch sites have relatively high success rates.



Distances between a launch site to its proximities

Shows a selected launch site to its proximities such as railway, highway, coastline, with distance calculated are displayed

Predictive analysis (Classification)

Machine Learning algorithms were applied on the data to predict if the rocket landed or not, the algorithms such as decision tree, SVM, KNN and Logistic regression algorithms were used

Classification Accuracy

As displayed in the form of a table, all the accuracies of the data on the test data set, the accuracy of the Decision Tree algorithm is the highest with 88.9% and KNN and Support Vector Machine algorithms having next best accuracies if 84.8%.

	Algorithms	Accuracies
0	DecisionTree	0.889286
1	KNN	0.848214
2	SVM	0.848214
3	LogisticReg	0.846429

Confusion Matrix

The confusion matrix displayed clearly shows the actual and predicted outcomes that landed, and which did not land such that out of 18 cases, 12 missions did land and the algorithm correctly predicted and 3 missions did not land while the algorithm predicted it to be landed.