IBM DATA SCIENCE CAPSTONE PROJECT

Provided by: Sofia Shahreki Mojahed

GitHub Link: https://github.com/sofiamoj/Space-X-Falcon-9

OUTLINE

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

EXECUTIVE SUMMARY

Summary of methodologies

- Data Collection via API, Web Scraping
- Exploratory Data Analysis (EDA) with Data Visualization
- EDA with SQL
- Interactive Map with Folium
- Dashboards with Plotly Dash
- Predictive Analysis

Summary of all results

- Exploratory Data Analysis results
- Interactive maps and dashboard
- Predictive results

Project background and context:

- This project delves into predicting the success of Falcon 9 first stage landings, a key factor in SpaceX's cost-effectiveness. SpaceX boasts a launch cost of $62 million compared to competitors exceeding $165 million. This significant difference hinges on their first-stage reusability. By forecasting landing outcomes, we can estimate launch costs, valuable information for potential competitors seeking to challenge SpaceX's dominance.

Problems we want to find answers:

- What are the main characteristics of a successful or failed landing ?
- What are the effects of each relationship of the rocket variables on the success or failure of a

  landing ?
- What are the conditions which will allow SpaceX to achieve the best landing success rate ?

INTRODUCTION

METHODOLOGY

- **Data collection methodology:**
  - SpaceX REST API
  - Web Scrapping from Wikipedia
- **Perform data wrangling**
  - Dropping unnecessary columns
  - One Hot Encoding for classification models
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
  - Classification Model Selection
  - Model Tuning and Evaluation

DATA COLLECTION

SpaceX API

- This dataset was collected from the SpaceX REST API.
- The API provides data about SpaceX launches, including rocket details, payloads, launch specifications, landing intentions, and outcomes.
- Our goal is to predict the likelihood of SpaceX attempting a rocket landing based on this launch data.
- The SpaceX REST API uses endpoints (URLs) starting with api.spacexdata.com/v4/ for data access.

Use SpaceX REST API

API Returns SpaceX data in .JSON

Normalize data into flat data file such as .csv

Web Scrapping

- While web scraping Wikipedia with BeautifulSoup is another method for obtaining Falcon 9 launch data, the SpaceX REST API offers a more structured and up-to-date source.

Get HTML Response from Wikipedia

Extract data using beautiful soup

Normalize data into flat data file such as .csv

DATA COLLECTION: SPACEX API

1. Getting Response from API
2. Convert Response to JSON File
3. Transform data
4. Create dictionary with data
5. Create dataframe
6. Filter dataframe
7. Export to file

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url)
```

```
data = response.json()
data = pd.json_normalize(data)
```

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
getBoosterVersion(data)
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

```
data = pd.DataFrame.from_dict(launch_dict)
```

```
data_falcon9 = data[data['BoosterVersion']!='Falcon 1']
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

DATA COLLECTION: SCRAPING

1. Getting Response from HTML → 2. Create Beautiful Soup Object → 3. Find all tables → 4. Get column names → 5. Create dictionary → 6. Add data to keys → 7. Create dataframe from dictionary → 8. Export to file

```
response = requests.get(static_url)
```

```
soup = BeautifulSoup(response.text, "html5lib")
```

```
html_tables = soup.findAll('table')
```

```
for th in first_launch_table.find_all('th'):
    name = extract_column_from_header(th)
    if name is not None and len(name) > 0 :
        column_names.append(name)
```

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

```
extracted_row = 0
#Extract each table
for table_number,table in enumerate(soup.find_all
   # get table row
    for rows in table.find_all("tr"):
        #check to see if first table heading is a
        if rows.th:
            if rows.th.string:
                flight_number=rows.th.string.stri
                flag=flight_number.isdigit()
```

See notebook for the rest of code

```
df=pd.DataFrame(launch_dict)
```

```
df.to_csv('spacex_web_scraped.csv', index=False)
```

- **Landing Outcomes in SpaceX Launch Data**
- Landing attempts can have various success rates.
- **Successful Landing:**
  - True Ocean: Successful landing on a designated ocean area.
  - True RTLS: Successful landing on a designated ground pad.
  - True ASDS: Successful landing on a drone ship.
- **Unsuccessful Landing:**
  - False Ocean: Attempted landing at sea, but failed.
  - False RTLS: Attempted landing on ground pad, but failed.
  - False ASDS: Attempted landing on drone ship, but failed.
- **Training Label:**
  - 1: Successful Landing (Any of the "True" outcomes)
  - 0: Unsuccessful Landing (Any of the "False" outcomes)

Each launch aims to a dedicated orbit, and here are some common orbit types:

1. Perform Exploratory Data Analysis EDA on dataset
2. Calculate the number oflaunches at each site
3. Calculate the number and occurrence of each orbit
4. Calculate the number and occurrence of mission outcome per orbit type
5. Export dataset as .CSV
6. Create a landing outcome label from Outcome column
7. Work out success rate for every landing in dataset

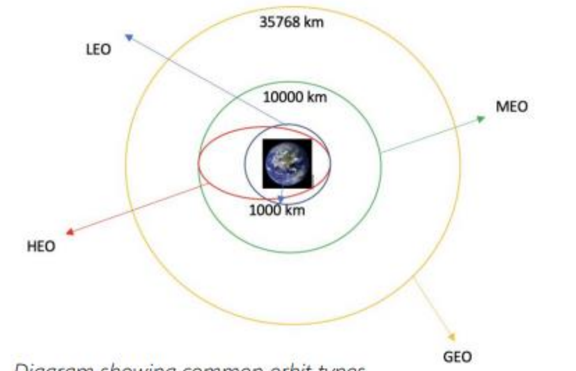


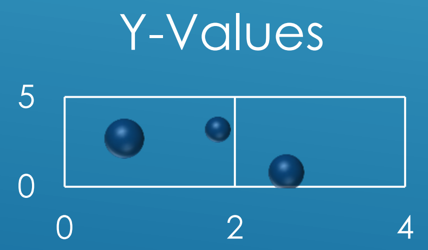
Diagram showing common orbit types SpaceX uses

DATA WRANGLING

EDA WITH DATA VISUALIZATION

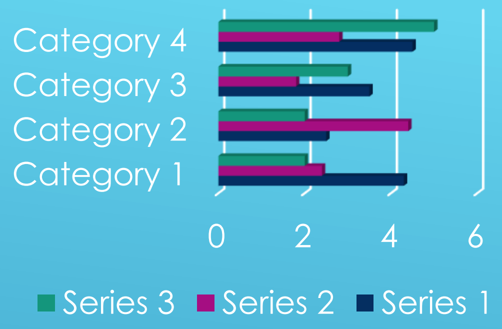
Scatter Graphs

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Orbit vs. Flight Number
- Payload vs. Orbit Type
- Orbit vs. Payload Mass



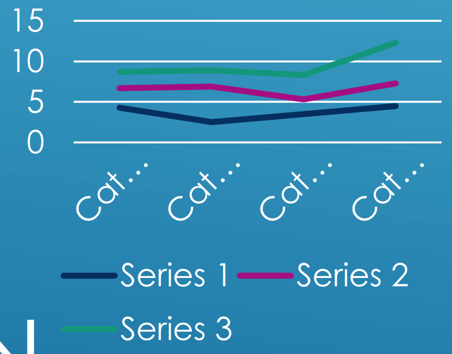
Bar Graph

- Success rate vs. Orbit



Line Graph

- Success rate vs. Orbit



EDA WITH SQL

- **Exploring Launch Locations:**
  - **Unique Launch Sites:** Displayed all unique launch site names used for SpaceX missions.
  - **Cape Canaveral Focus:** Listed the top 5 launches originating from launch sites beginning with "CCA" (likely Cape Canaveral).
- **Payload Analysis:**
  - **NASA's Contribution:** Calculated the total payload mass lifted by boosters launched for NASA's CRS program.
  - **F9 v1.1 Performance:** Displayed the average payload mass carried by the F9 v1.1 booster version.
- **Landing Outcomes:**
  - **Ground Pad Success:** Identified the date of the first successful ground pad landing.
  - **Drone Ship Mastery:** Listed boosters with successful drone ship landings and payload mass between 4,000 and 6,000 kg.
- **Mission Performance Summary:**
  - **Success vs. Failure:** Counted the total number of successful and failed missions.
  - **Heavy Lifters:** Identified the booster versions that carried the heaviest payloads.
- **Detailed Analysis (2015):**
  - **Monthly Drone Ship Failures:** Presented a table displaying month names, failed drone ship landings, booster versions, and launch sites for all launches in 2015.
- **Landing Success Trends (2010-2017):**
  - **Ranked Success Rates:** Ranked launch dates between 04-06-2010 and 20-03-2017 by the number of successful landings (descending order).

BUILD AN INTERACTIVE MAP WITH FOLIUM

- **Map Components:**
  - **Center Stage:** The map centers on NASA's Johnson Space Center in Houston, Texas.
  - **Launch Sites Highlighted:** Red circles mark each launch site location, with labels displaying their names.
  - **Clustering Close Calls:** When multiple sites share coordinates, a cluster groups them for better visualization.
  - **Landing Success at a Glance:** Green markers indicate successful landings, while red markers signify failures.
  - **Connecting the Dots:** Lines connect launch sites to nearby points of interest (railways, highways, coastlines, cities) to understand their accessibility.
- **Benefits:**
  - **Clear Overview:** Easily visualize all launch sites and their surroundings.
  - **Landing Performance:** Quickly identify successful and unsuccessful landing locations.
  - **Contextual Understanding:** Map infrastructure elements can reveal potential factors influencing launch decisions.

- **Exploring Launch Data with Dash:**

- **Site Selection Flexibility:** A dropdown menu allows users to choose a specific launch site or analyze all sites combined.
- **Success at a Glance:** A pie chart visually depicts the overall success rate for the chosen launch site, dividing launches into successful and failed categories.
- **Zooming in on Payload Mass:** A range slider empowers users to focus on a specific range of payload mass, allowing for a more granular analysis.
- **Success vs. Payload Mass:** A scatter plot reveals the relationship between launch success and payload mass. This visualization can help identify potential trends or correlations between these factors.

BUILD A DASHBOARD WITH PLOTLY DAS

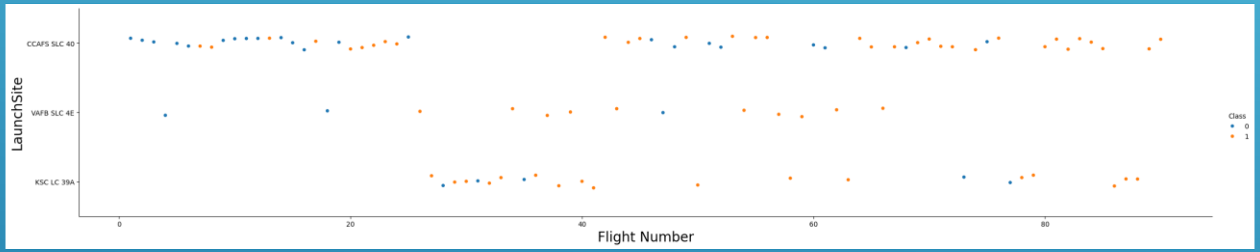
- **Data Preparation:**
  - **Fueling the Model:** We start by loading the dataset containing SpaceX launch data.
  - **Balancing the Scales:** Data normalization ensures all features are on a similar scale, improving model performance.
  - **Training vs. Testing:** The data is split into training and testing sets. The training set teaches the model, while the testing set evaluates its generalizability.
- **Model Selection and Training:**
  - **Picking the Right Tools:** We explore different machine learning algorithms to find the best fit for predicting landing outcomes.
  - **Fine-Tuning the Approach:** GridSearchCV helps us efficiently test various parameter combinations for each algorithm.
  - **Training Champions:** The training dataset is used to train models with the best hyperparameter configurations identified by GridSearchCV.
- **Model Evaluation:**
  - **Finding the Sweet Spot:** We evaluate each model's performance using the testing dataset and identify the model with the highest accuracy.
  - **Beyond Accuracy:** Confusion matrices provide a deeper understanding of model performance, revealing potential areas for improvement.
- **Choosing the Champion:**
  - **Accuracy Wins:** The model with the best overall accuracy on the testing set will be chosen as the final prediction model (detailed results available in the notebook).

PREDICTIVE ANALYSIS (CLASSIFICATION)

- **Unveiling SpaceX Launch Insights**
- **Key Analysis Areas:**
  - **Exploratory Data Analysis:** We delved into the data to uncover patterns and trends.
  - **Interactive Exploration**
  - **Predictive Modeling:** We built a model to predict SpaceX landing outcomes.

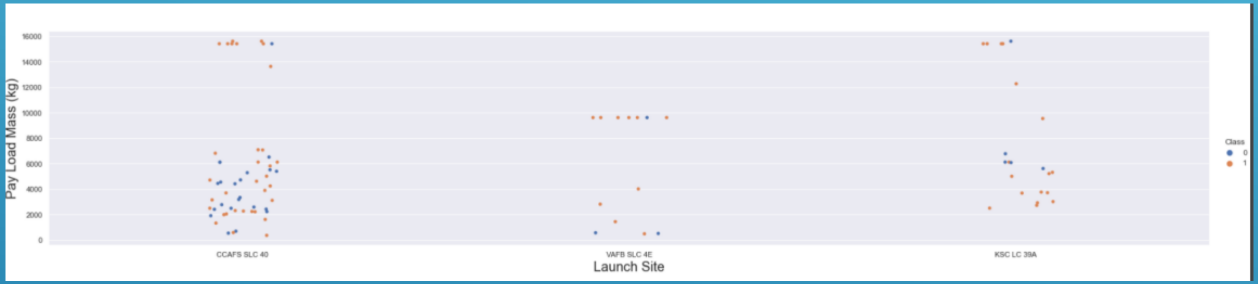
RESULTS

FLIGHT NUMBER VS. LAUNCH SITE



The more amount of flights at a launch site the greater the success rate at a launch site.

PAYLOAD MASS VS. LAUNCH SITE

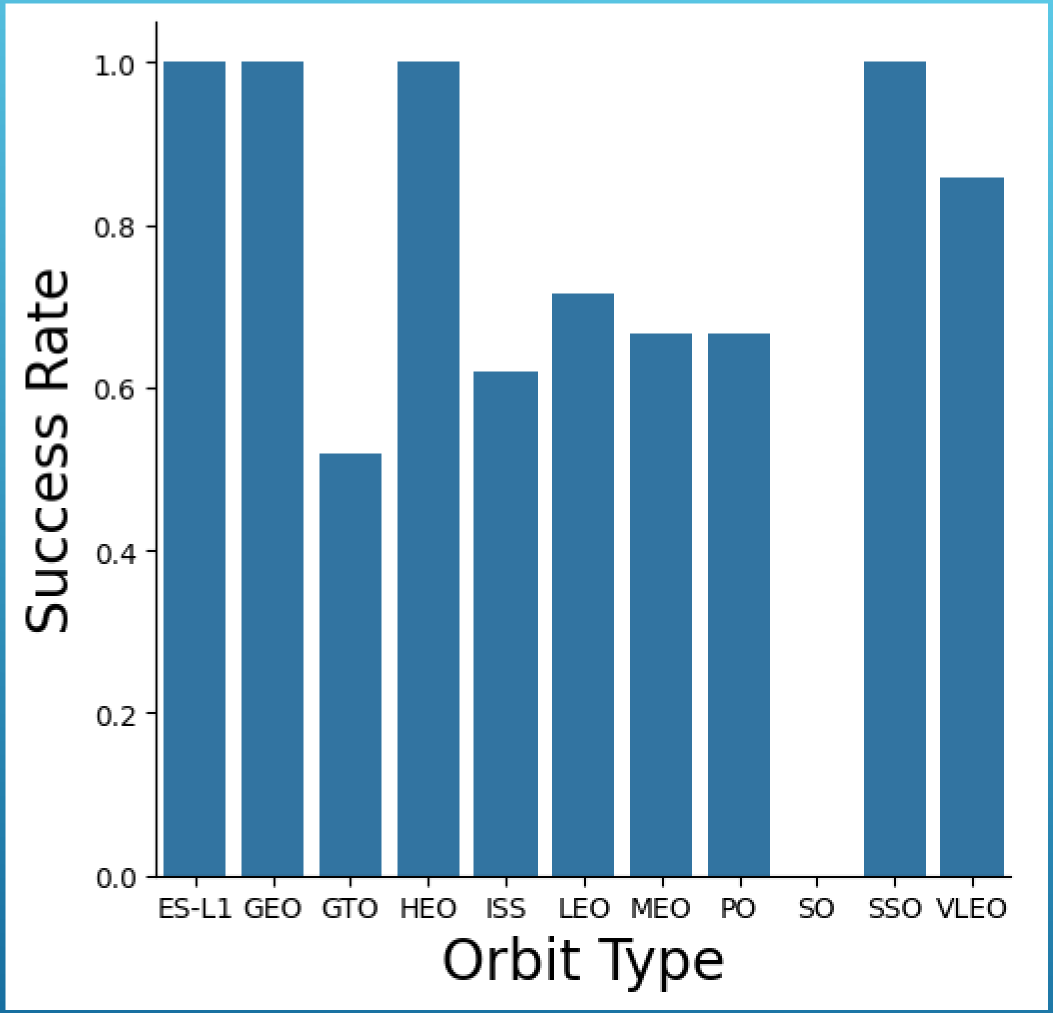


Launch Site vs. Payload Mass: Examining Success Rates at CCAFS SLC-40

Finding: Our analysis revealed no clear correlation between payload mass and launch success rate at Cape Canaveral's Launch Site SLC-40 (CCAFS SLC-40).

Explanation: While the visualization might have hinted at a possible trend, the data doesn't provide a statistically significant link between heavier payloads and higher success rates for rockets launched from CCAFS SLC-40.

Implications: This suggests other factors likely play a more significant role in launch success at this particular launch site. Further investigation into these other factors is necessary to determine the key drivers of successful launches at CCAFS SLC-40.

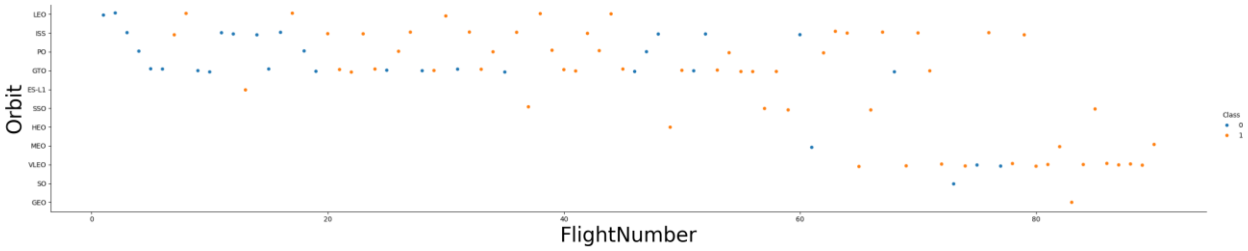


SUCCESS RATE VS. ORBIT TYPE

Orbit GEO,HEO,SSO,ES-L1 has the best Success

Rate

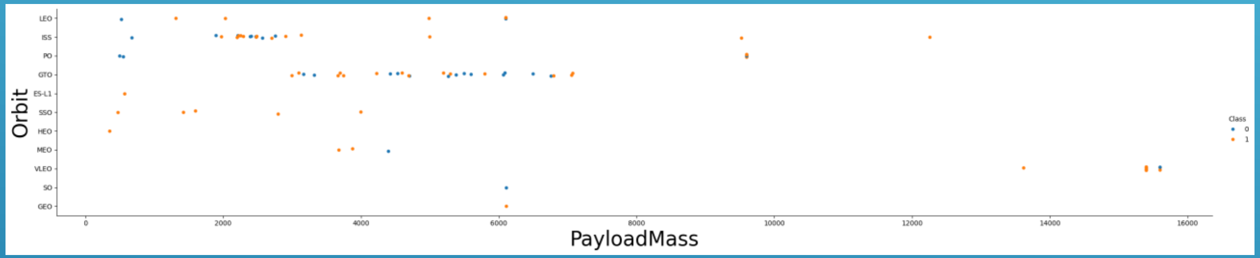
FLIGHT NUMBER VS. ORBIT TYPE



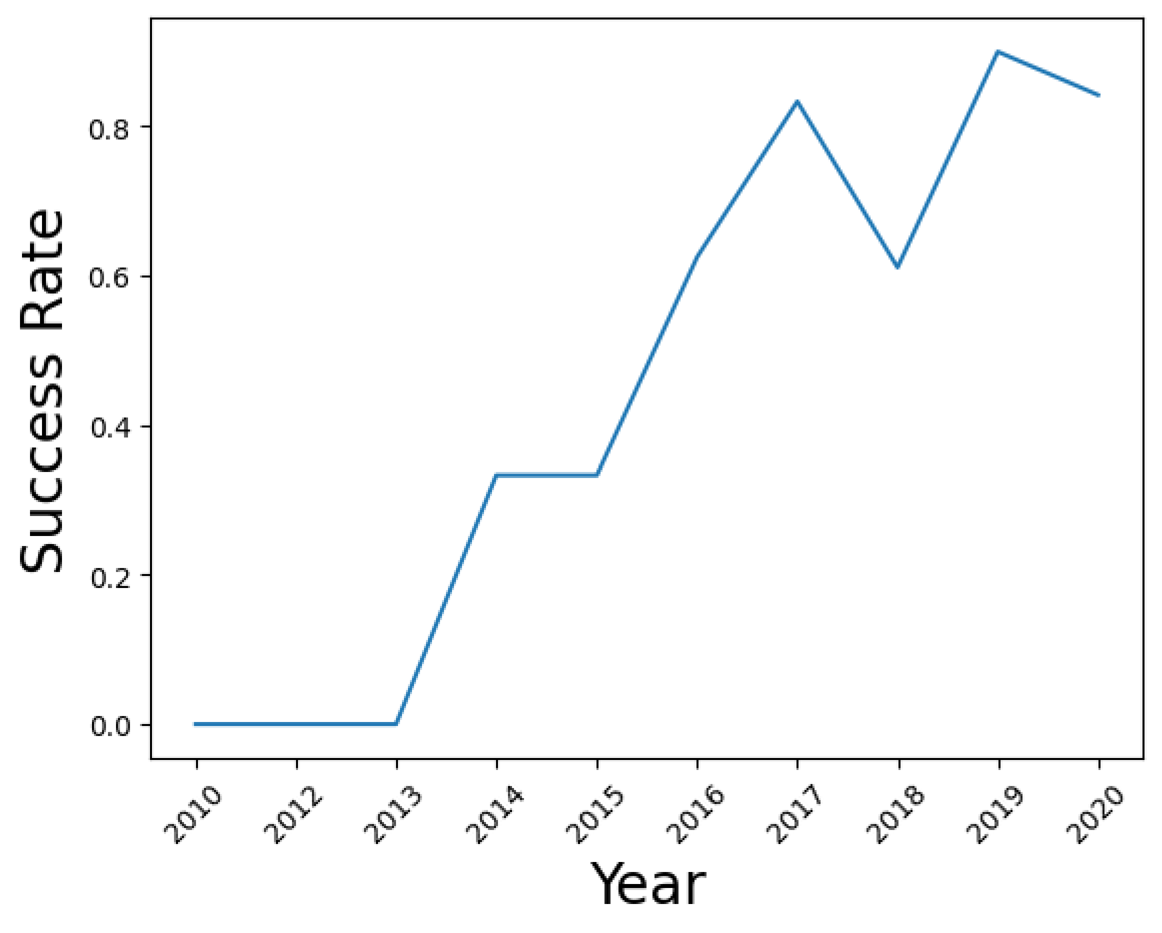
There appears to be no significant correlation between the number of flights and success rate for GTO orbits.

The data points seem scattered, suggesting other factors may play a more significant role in the success of GTO launches.

PAYLOAD VS. ORBIT TYPE



You should observe that Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.



LAUNCH SUCCESS YEARLY TREND

you can observe that the success rate since 2013 kept increasing till 2020

- **Launch_Site**

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

```
%sql select distinct launch_site from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

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

SQL QUERY

- **Launch site names begin with `CCA`:**
- select TOP 5 * from tblSpaceX WHERE Launch_Site LIKE 'KSC%'

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

SQL QUERY

- **Total payload mass 27 Total Payload Mass by Customer NASA (CRS)**

total_payload_mass
45596

- **Average Payload Mass carried by booster version F9 v1.1**

average_payload_mass
2534.6666666666665

SQL QUERY

- **The successful landing outcome in the drone ship was achieved**

first_successful_landing
2015-12-22

SQL QUERY

- **The names of the boosters which have success in drone ship and have 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

- The total number of successful and failed mission outcomes

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

- The names of the booster versions which have carried the maximum payload mass. Use a subquery

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

SQL QUERY

- **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	Date	Landing_Outcome	Booster_Version	Launch_Site
01	2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

SQL QUERY

- **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.**

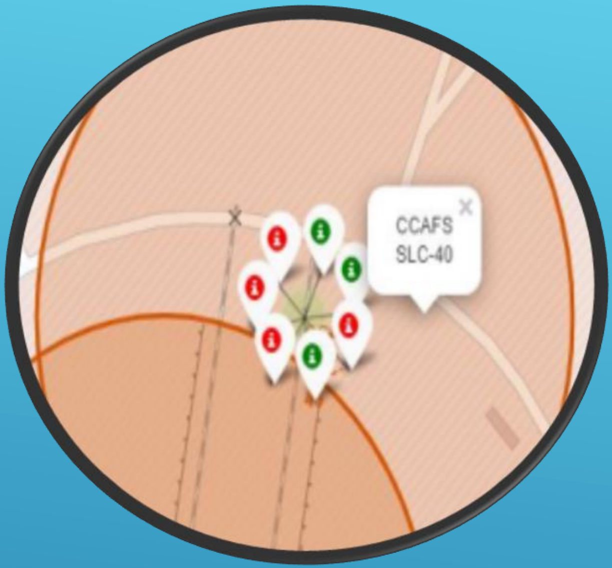
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

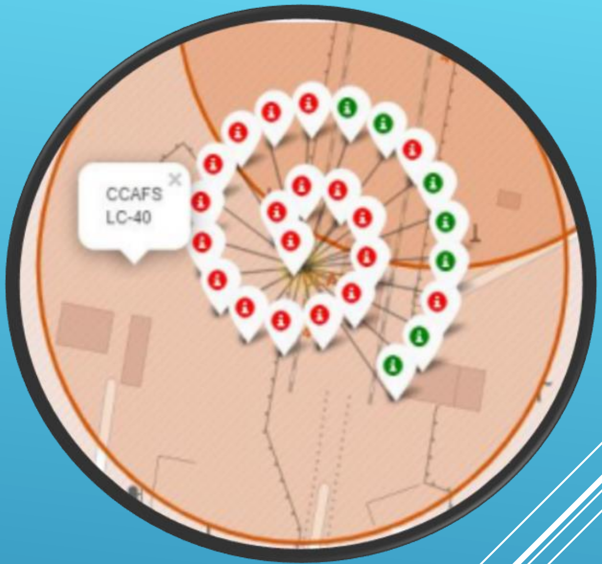
INTERACTIVE VISUAL ANALYTICS WITH FOLIUM

SQL QUERY

FOLIUM MAP : GROUND STATIONS

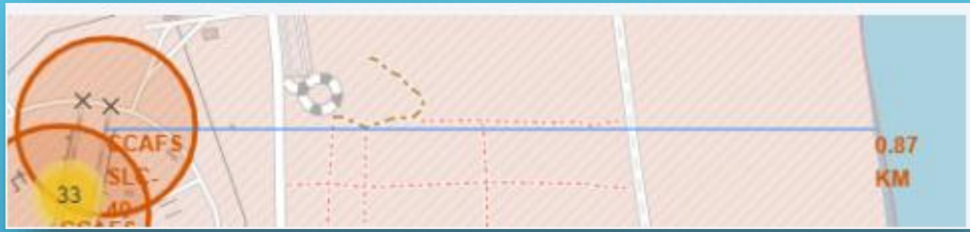
- The Space X launch sites are located on the coast of the United States

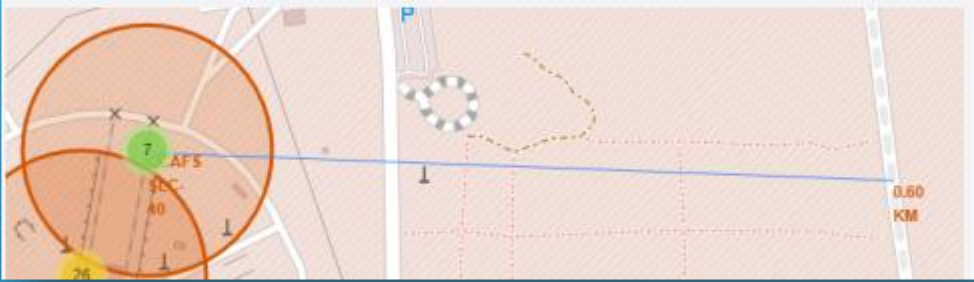


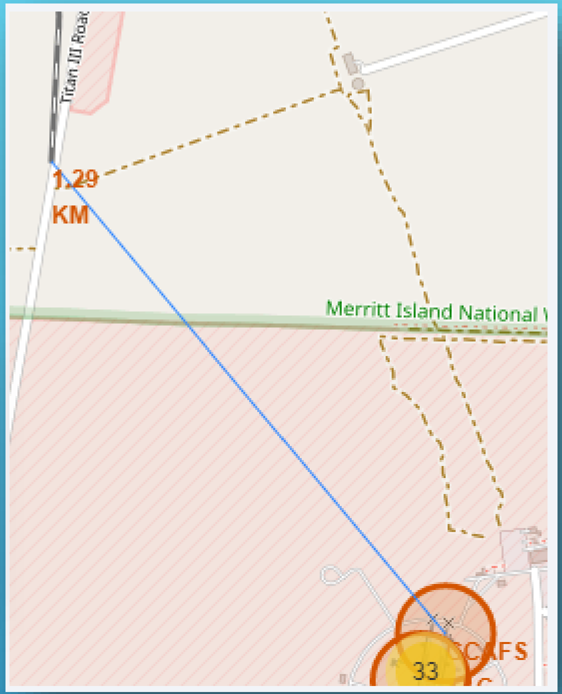


COLOR LABELED MARKERS

The green marker represents successful launches. The red marker represents unsuccessful launches. Note that KSC LC-39A has a higher launch success rate





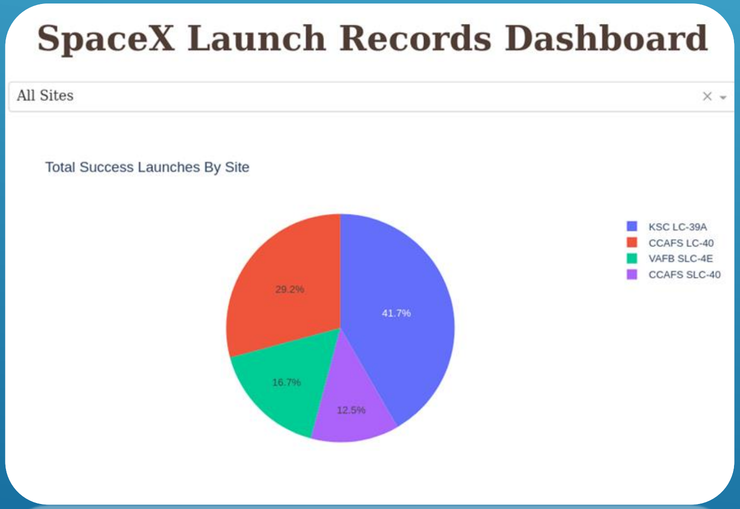


DISTANCES BETWEEN CCAFS SLC-40 AND ITS PROXIMITIES

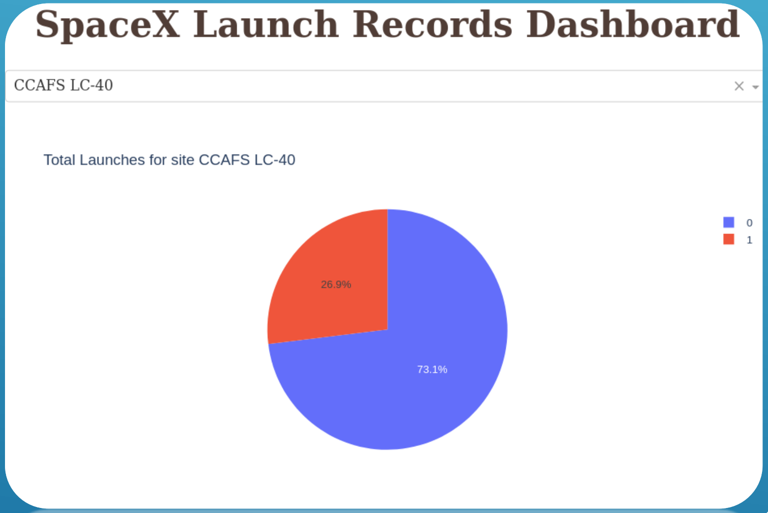
The CCAFS SLC-40 is in close proximity to railways
The CCAFS SLC-40 Is in close proximity to highways
The CCAFS SLC-40 Is in close proximity to coastline
CCAFS SLC-40 does not keep certain distance away from cities

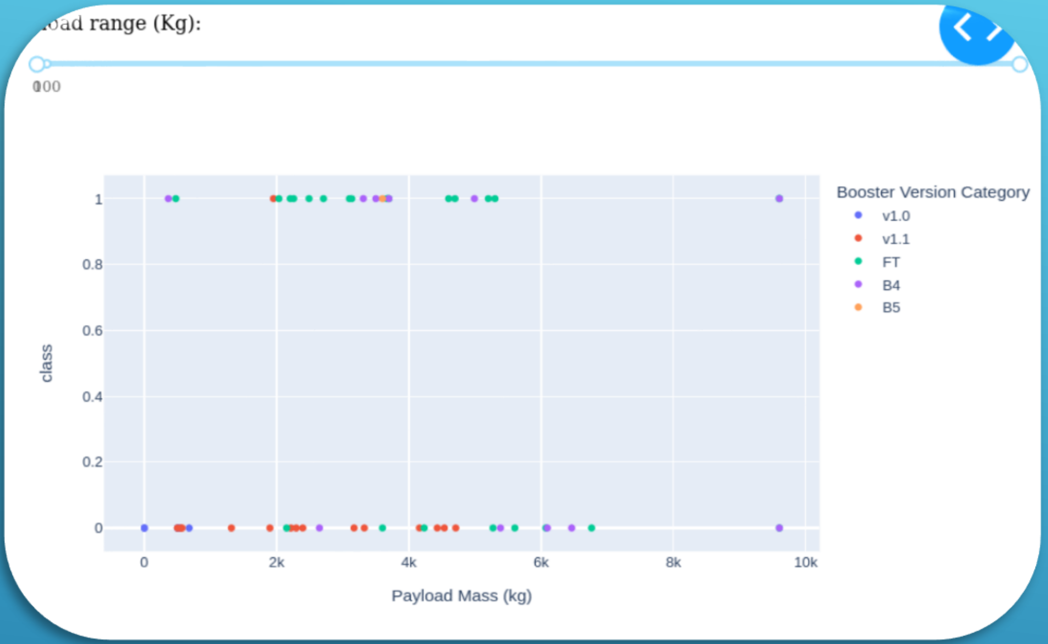
DASHBOARD WITH PLOTLY DASH

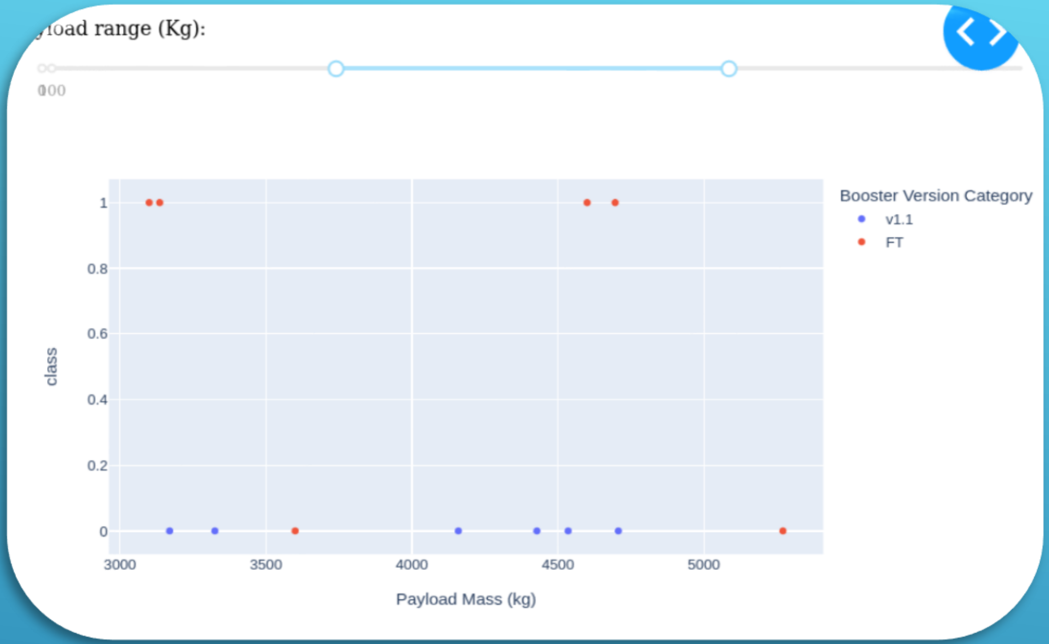
- **Total success by Site**
- The KSC LC-39A has the best success rate of launches



- **Total success launches for Site KSC LC-39A**
- The KSC LC-39A has achieved a 76.9% success rate while getting a 23.1% failure rate.







PAYLOAD MASS VS OUTCOME FOR ALL SITES WITH DIFFERENT PAYLOAD MASS SELECTED

Low weighted payloads have a better success rate than the heavy weighted payloads.

PREDICTIVE ANALYSIS (CLASSIFICATION)

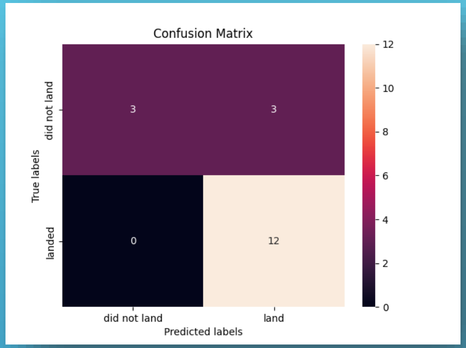
	ML Method	Accuracy Score (%)
0	Support Vector Machine	83.333333
1	Logistic Regression	83.333333
2	K Nearest Neighbour	83.333333
3	Decision Tree	83.333333

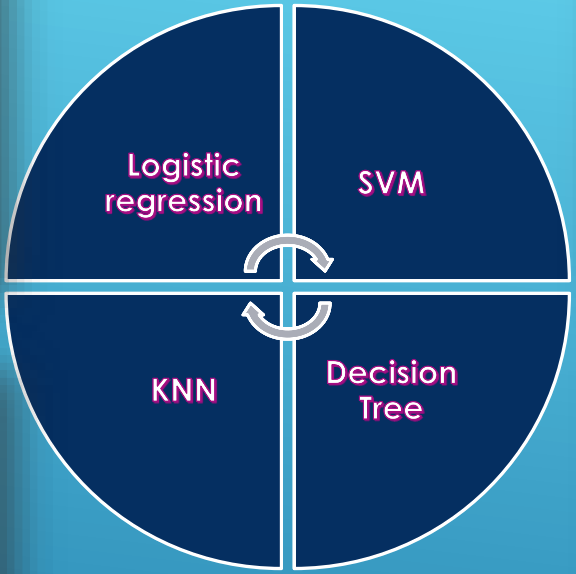
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.666667	0.800000
F1_Score	0.888889	0.888889	0.800000	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

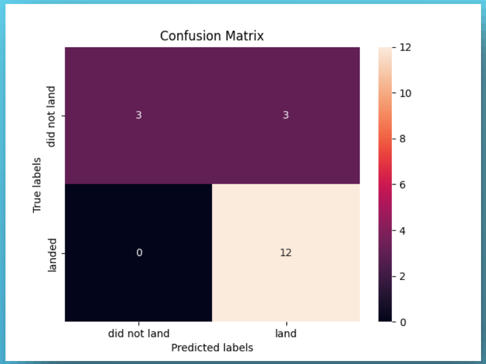
```
Best model is DecisionTree with a score of 0.8767857142857143
Best params is : {'criterion': 'gini', 'max_depth': 14, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 1
0, 'splitter': 'random'}
```

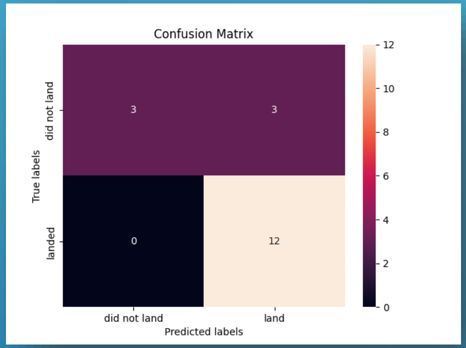
FIND THE METHOD PERFORMS BEST

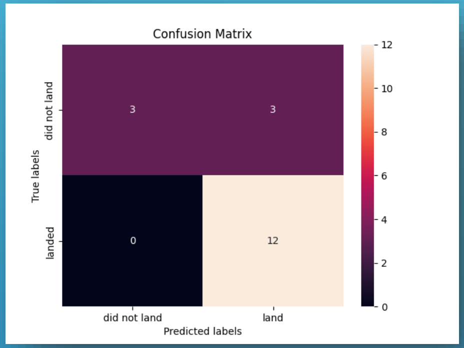
The best model is the Decision Tree with a score of 0.8767857142857143











CONFUSION MATRIX

As the test accuracy are all equal, the confusion matrices are also identical. The main problem of these models is false positives.

		Actual values	
		1	0
Predicted values	1	TP	FP
	0	FN	TN

CONCLUSIONS

Let's discuss the key factors that contribute to the success of space missions:

- Launch Site Selection:

The choice of launch site significantly impacts mission success. Factors such as geographical location, weather conditions, and infrastructure play a crucial role. For instance, Kennedy Space Center's Launch Complex 39A (KSC LC-39A) has consistently demonstrated better performance. However, understanding why certain sites excel requires additional data, such as atmospheric conditions or historical success rates.

- Orbit Considerations:

Different orbits have varying success rates. Some of the most successful orbits include

- ✓ Geostationary Earth Orbit (GEO)
- ✓ Highly Elliptical Orbit (HEO)
- ✓ Sun-Synchronous Orbit (SSO)
- ✓ Earth-Sun Lagrange Point 1 (ES-L1)

- Payload mass is also a critical consideration. Some orbits favor lighter payloads, while others accommodate heavier ones.

- Knowledge Accumulation and Learning:

Over time, knowledge gained from previous launches contributes to improved success rates. Lessons learned from failures lead to better mission planning and execution. Data science techniques help analyze historical data, identify patterns, and extract insights. Machine learning models can predict outcomes based on these patterns.

- Decision Tree Algorithm:

In the dataset, the Decision Tree Algorithm emerged as the best model, even though test accuracy was identical across various models. Decision trees are interpretable and allow us to visualize decision rules. They split data based on features, creating a tree-like structure. The choice of this algorithm likely stems from its better training accuracy, which indicates effective learning from historical data


THANKS FOR YOUR ATTENTION