SpaceX Analysis

IBM Data Science Capstone Project

Hazem Haffouz

Executive Summary

Problem Statement:

- "Understanding SpaceX First Stage Landing Success ration."

Methodologies Employed:

- Data Collection: Utilized SpaceX API and web scraping techniques to gather comprehensive datasets.
- Data Wrangling: Leveraged Python libraries such as Pandas and NumPy for data cleaning and transformation.
- Exploratory Data Analysis (EDA): Conducted in-depth analysis using Python for numerical data and SQL for database queries.
- Data Visualization: Employed Matplotlib for basic visualizations, and advanced tools like Plotly and Folium for interactive charts and maps.
- Predictive Modeling: Utilized classification algorithms to predict launch success rates.

Introduction: Predicting the Success of SpaceX Falcon 9 First Stage Landings

Background

The commercial space industry has been revolutionized by the advent of reusable rocket technology, primarily driven by SpaceX's Falcon 9 rocket. The company's ambitious goal to make space travel more affordable hinges on the reusability of its rocket stages, particularly the first stage, which constitutes a significant portion of the rocket's total cost. By landing the first stage back on Earth successfully, SpaceX aims to refurbish and reuse it, thereby dramatically reducing the cost per launch.

Importance

Understanding the factors that influence the successful landing of Falcon 9's first stage is of paramount importance, not just for SpaceX, but for the entire aerospace industry. It offers insights into mission planning, risk assessment, and ultimately, cost-saving measures that can make or break the economic viability of space missions. For competing firms and contractors who might want to bid against SpaceX for launching services, knowing the likelihood of a successful landing could be a crucial variable in their cost models.

Introduction: Predicting the Success of SpaceX Falcon 9 First Stage Landings

Objective

The primary objective of this project is to develop predictive models that can accurately forecast the outcome of a Falcon 9 first stage landing based on various mission parameters. We seek to answer the question: given specific conditions and configurations, will the first stage successfully land back on Earth?

Methodology

To achieve this objective, we utilize a dataset that includes details about past SpaceX launches, such as payload mass, orbit type, launch site, and numerous other variables. We employ several machine learning algorithms, including Logistic Regression, Support Vector Machines, Decision Trees, and K-Nearest Neighbors, to build and optimize predictive models. These models are then rigorously evaluated on a test set to ascertain their accuracy and reliability.

Introduction: Predicting the Success of SpaceX Falcon 9 First Stage Landings

Scope

This project covers the end-to-end process of predictive modeling, from initial data exploration and preprocessing to model training, hyperparameter tuning, and evaluation. We also discuss the practical implications of our findings and propose directions for future research in this field.

Organization

The report is organized as follows:

- Data Collection and Preprocessing
- Exploratory Data Analysis (EDA)
- Feature Engineering and Standardization
- Model Building and Hyperparameter Tuning
- Evaluation and Comparison of Models
- Final Insights and Recommendations

Data Collection and Wrangling

SpaceX API:
The process that was involved in acquiring data was using python to pull data thorough SpaceX API, after that the data was transformed and loaded in dataframes, a lot of the data are IDs, So we used API again to Identify the columns we need,
Next we Filter out All data that is not relevant to Falcon 9.
Then we check for nulls, found around 31 null variable, to solve this, we used the mean as a replacement for those nulls.

Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia

Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia:
Another source of data using Python with `BeautifulSoup`
collect Falcon 9 historical launch records from a Wikipedia page titled `List of Falcon 9 and Falcon Heavy launches`
To be more precise, the launch records are stored in a HTML table,
Then the next step was parsing it and storing it in a dataframe.

Data Wrangling

Now, our objective is to find some patterns in the data and determine what would be the label for training supervised models.

For example, **`True`** `Ocean` means the mission outcome was successfully landed to a specific region of the ocean while **`False`** `Ocean` means the mission outcome was unsuccessfully landed to a specific region of the ocean. **`True RTLS`** means the mission outcome was successfully landed to a ground pad **`False RTLS`** means the mission outcome was unsuccessfully landed to a ground pad.**`True ASDS`** means the mission outcome was successfully landed on a drone ship **`False ASDS`** means the mission outcome was unsuccessfully landed on a drone ship.

Data Wrangling

So our Task will be exploring the data and Transforming it its suitable for our objective:

Finding the number of launches in each site

Calculating the number and occurrence of each orbit

Calculating the number and occurence of mission outcome of the orbits

Creating a landing outcome label from Outcome column

Determining the success rate of landing: which is 67%

Data Wrangling

Our key findings about the number of launches, Per Launch site, per orbit, per Outcome:

```
LaunchSite
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: count, dtype: int64
```

```
Outcome
True ASDS       41
None None       19
True RTLS       14
False ASDS       6
True Ocean       5
False Ocean      2
None ASDS        2
False RTLS       1
Name: count, dtype: int64
```

```
Orbit
GTO     27
ISS     21
VLEO    14
PO       9
LEO      7
SSO      5
MEO      3
ES-L1    1
HEO      1
SO       1
GEO      1
Name: count, dtype: int64
```

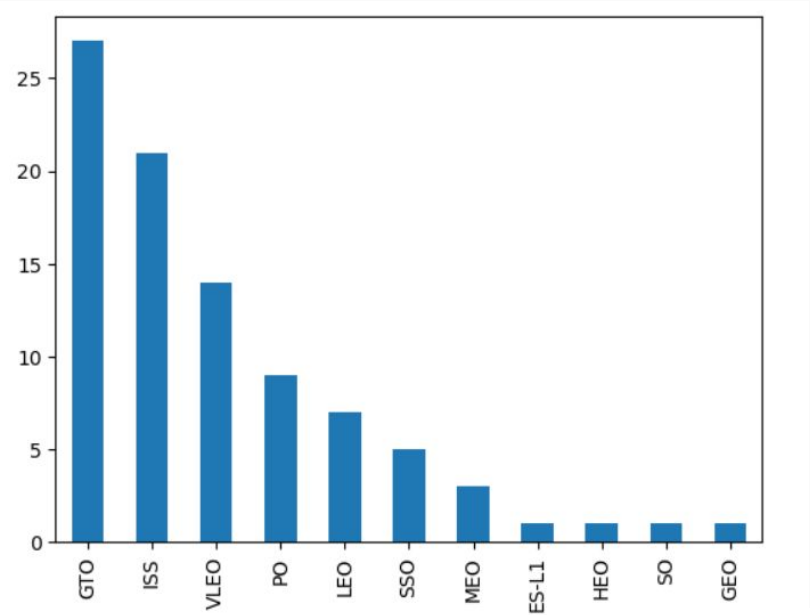
EDA and Visualization

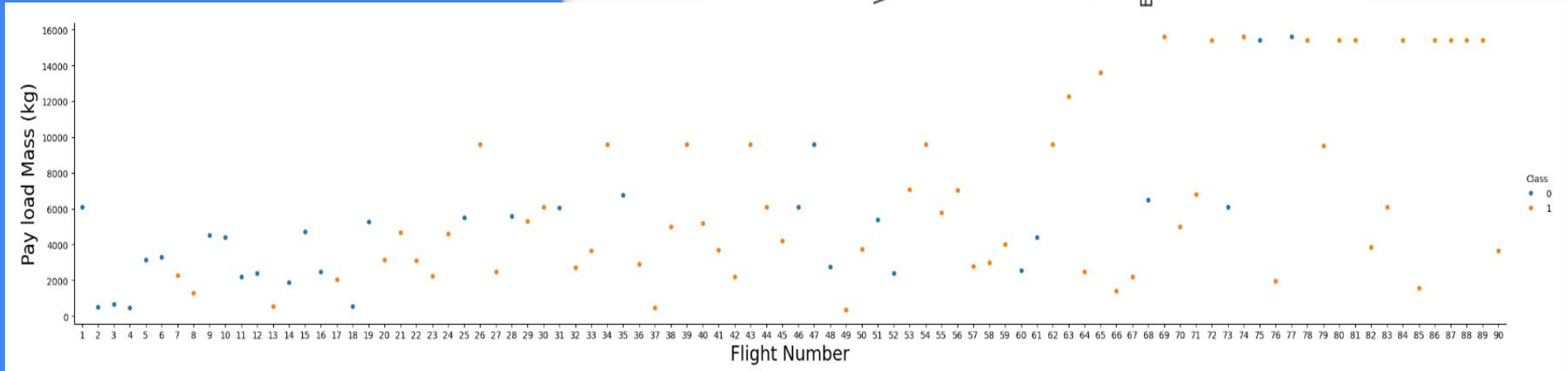
After we Verified and Transformed our data, we will predict if the Falcon 9 first stage will land successfully our Objective will be Perform exploratory Data Analysis and Feature Engineering using `Pandas` and `Matplotlib.`

EDA and Visualization

First, let's try to see how the `Flight Number` (indicating the continuous launch attempts.) and `Payload` variables would affect the launch outcome.

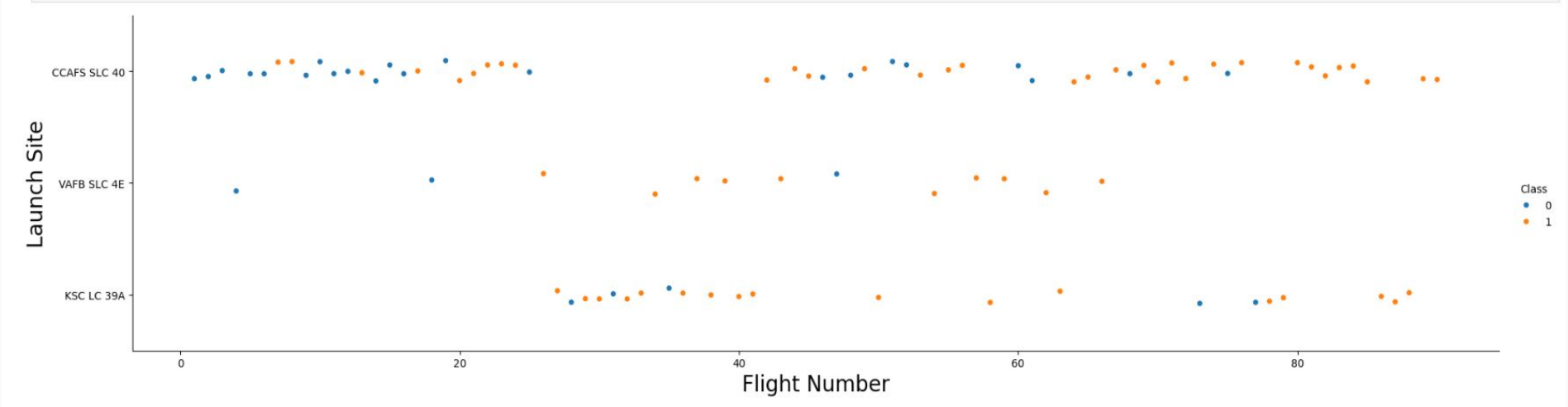
We see that different launch sites have different success rates. `CCAFS LC-40`, has a success rate of 60 %, while `KSC LC-39A` and `VAFB SLC 4E` has a success rate of 77%.





EDA and Visualization

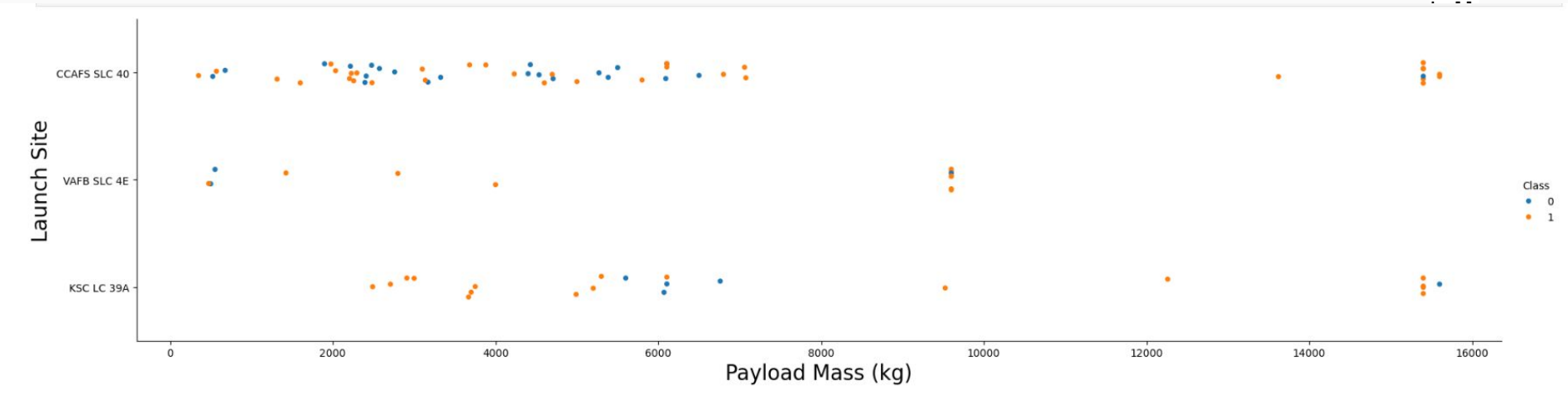
Next, let's drill down to each site visualize its detailed launch records.



EDA and Visualization

We also want to observe if there is any relationship between launch sites and their payload mass.

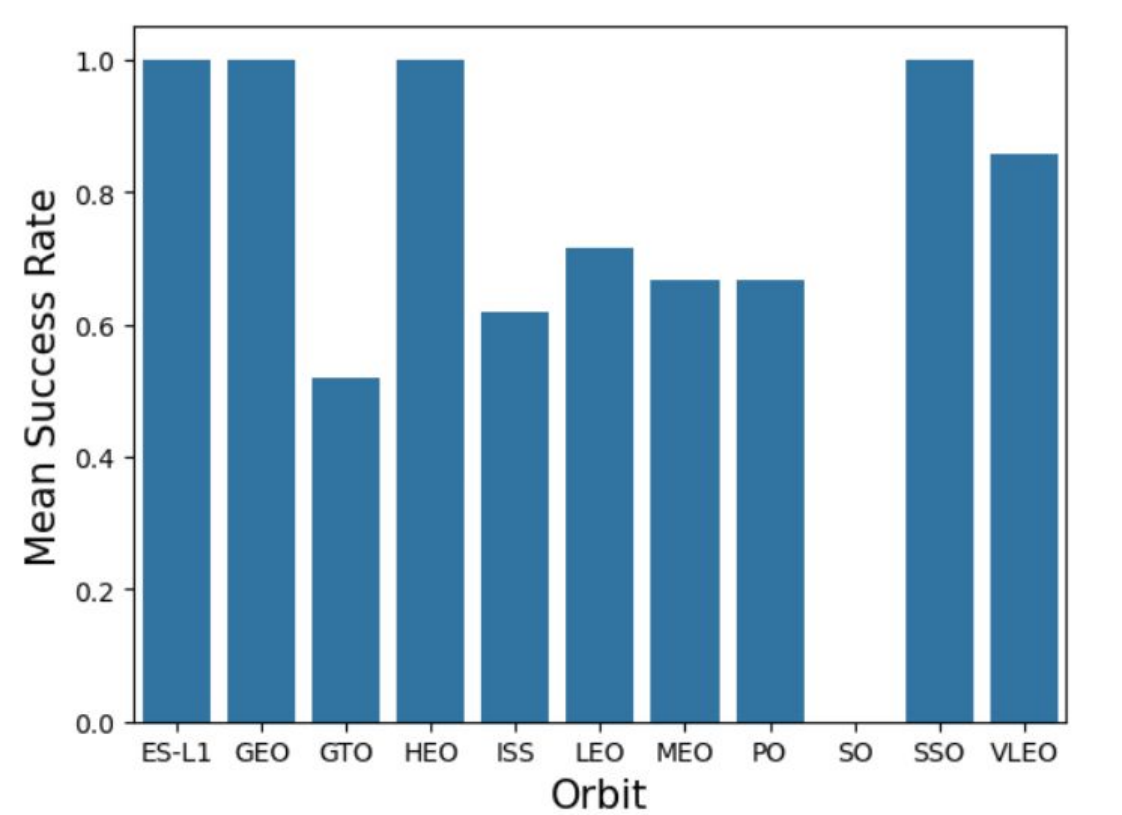
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000).



EDA and Visualization

Next, we want to visually check if there are any relationship between success rate and orbit type.

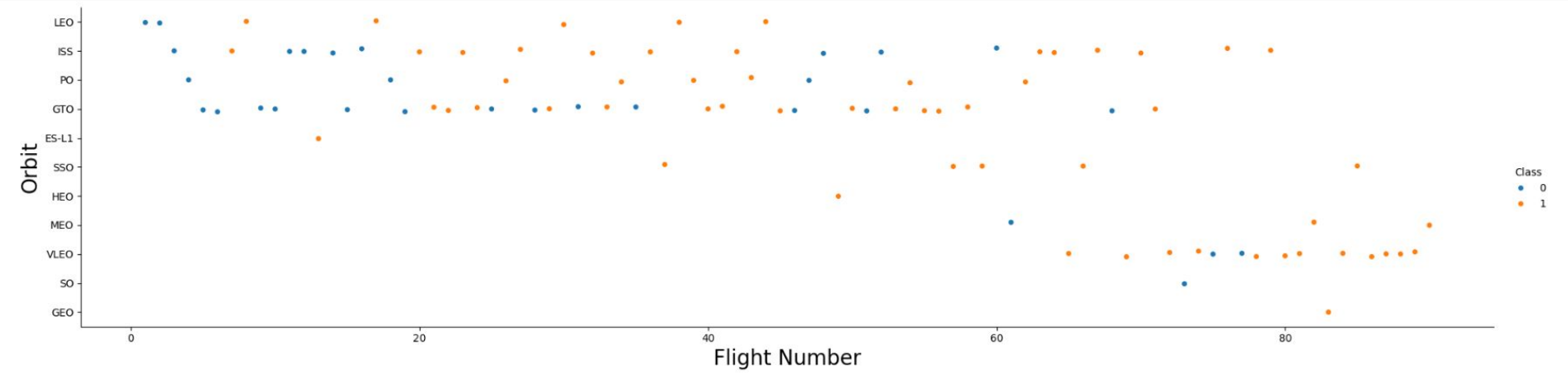
ES-L1 GEO HEO SSO have the highest success rate.



EDA and Visualization

For each orbit, we want to see if there is any relationship between FlightNumber and Orbit type.

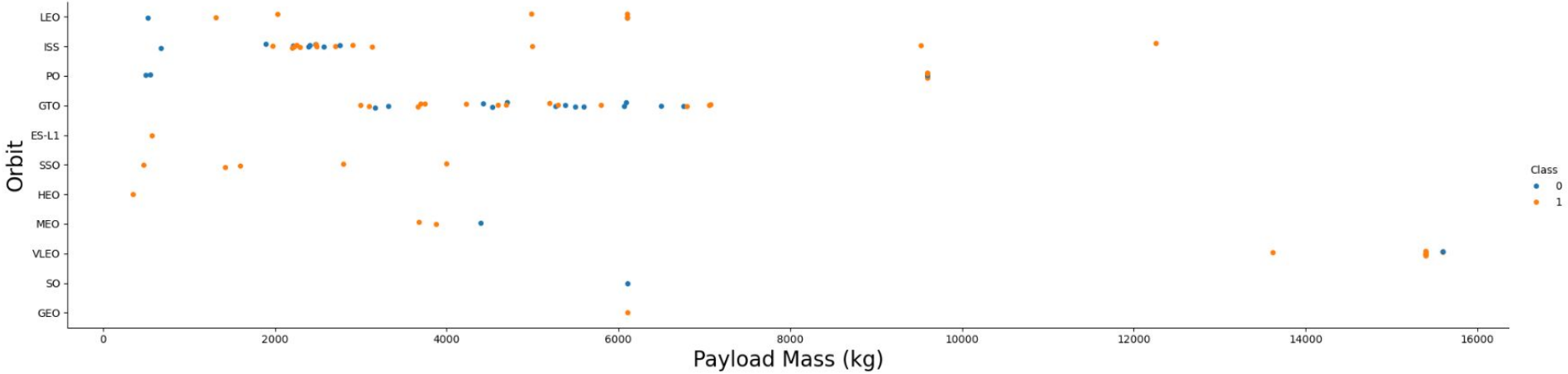
We s see that in the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



EDA and Visualization

Similarly, we can plot the Payload vs. Orbit scatter point charts to reveal the relationship between Payload and Orbit type

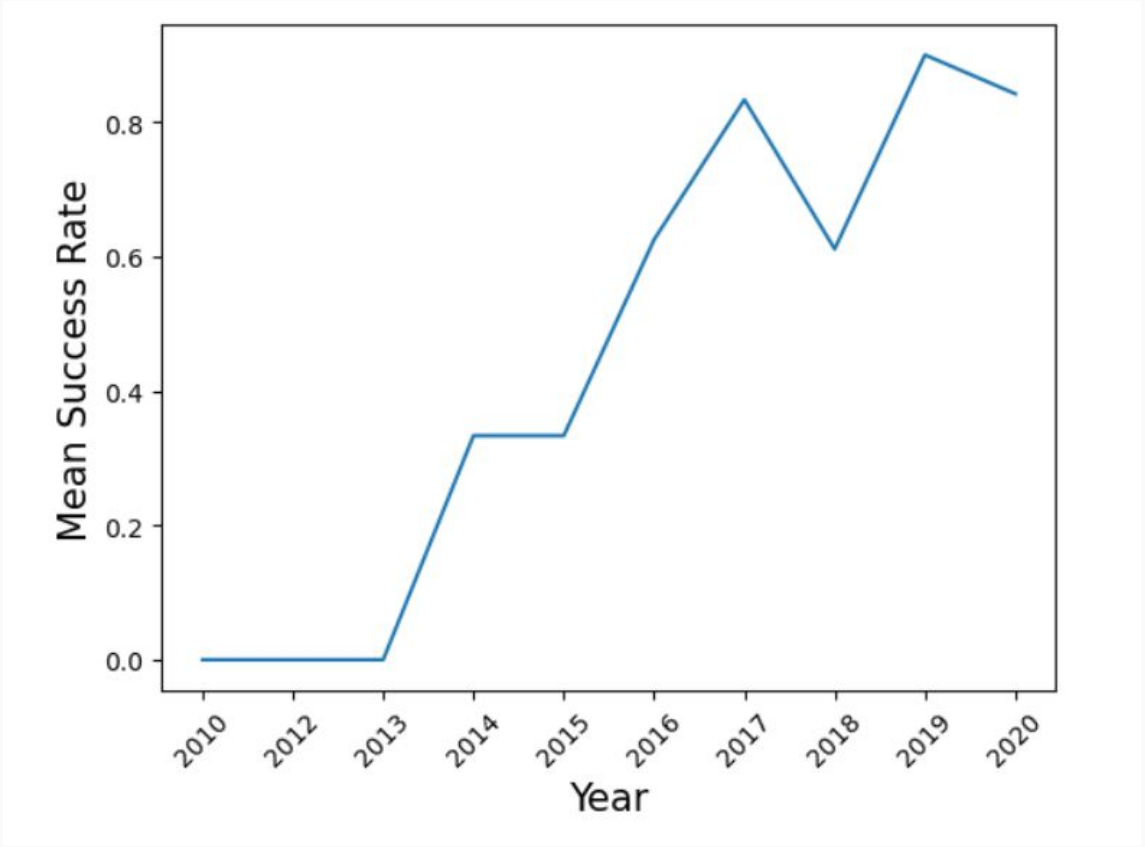
With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



EDA and Visualization

Visualize the launch success yearly trend

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



EDA and Visualization

By now, we obtained some preliminary insights about how each important variable would affect the success rate, we will select the features that will be used in success prediction in the future module.

	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

EDA with SQL

Understanding dataset includes a record for each payload carried during a SpaceX mission into outer space.

Lets Query the dataset to understand:

The names of the unique launch sites in the space mission

5 records where launch sites begin with the string 'CCA'

The total payload mass carried by boosters launched by NASA (CRS)

Average payload mass carried by booster version F9 v1.1

The date when the first successful landing outcome in ground pad was acheived.

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

The total number of successful and failure mission outcomes

The names of the booster_versions which have carried the maximum payload mass. Use a subquery

The records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

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.

Task 1

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

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE;
```

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

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

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("Payload_Mass__kg_") FROM SPACEXTABLE WHERE "Customer" LIKE 'NASA (CRS)%';
```

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

SUM("Payload_Mass__kg_")
48213

Display 5 records where launch sites begin with the string 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT AVG("Payload_Mass__kg_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1';
```

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

AVG("Payload_Mass__kg_")
2928.4

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing__Outcome" = 'Success (ground pad)';
```

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

MIN("Date")
None

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

```
%sql SELECT DISTINCT "Booster_Version" FROM SPACEXTABLE WHERE "Landing__Outcome" = 'Success (drone ship)' AND "Payload_Mass__kg_" BETWEEN 4001 AND 5999;
```

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

Booster_Version

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Payload_Mass__kg_" = (SELECT MAX("Payload_Mass__kg_") FROM SPACEXTABLE);
```

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

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 7

List the total number of successful and failure mission outcomes

```
%sql SELECT "Landing__Outcome", COUNT(*) FROM SPACEXTABLE GROUP BY "Landing__Outcome";
```

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

"Landing__Outcome"	COUNT(*)
Landing__Outcome	101

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr("Date", 6, 2) AS Month, "Landing__Outcome", "Booster_Version", "Launch_Site" FROM SPACEXTABLE WHERE "Landing__Outcome" LIKE 'Failure (drone ship)' AND substr("Date", 1, 4) = '2
```

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

Month	"Landing__Outcome"	Booster_Version	Launch_Site

Task 10

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.

```
%sql SELECT "Landing__Outcome", COUNT(*) AS Count FROM SPACEXTABLE WHERE "Date" BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY "Landing__Outcome" ORDER BY Count DESC;
```

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

"Landing__Outcome"	Count
Landing_Outcome	32

Analysis with Folium

The launch success rate may depend on many factors such as payload mass, orbit type, and so on. It may also depend on the location and proximities of a launch site, i.e., the initial position of rocket trajectories. Finding an optimal location for building a launch site certainly involves many factors and hopefully we could discover some of the factors by analyzing the existing launch site locations.

Our Objective will be, to Mark All Launch sites, Mark the Success and fails for each launch, Calculate the distance between launch site and key locations.

Analysis with Folium

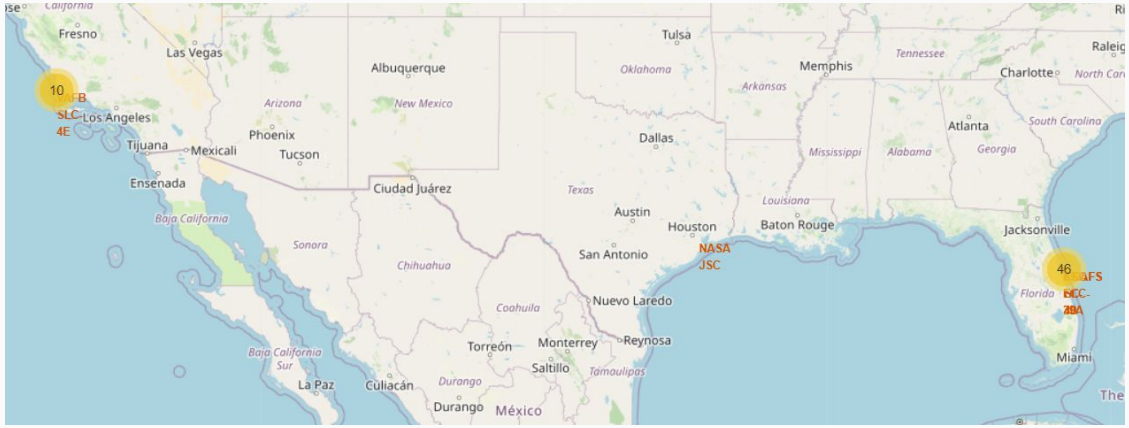
First, let's try to add each site's location on a map using site's latitude and longitude coordinate

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610745

Analysis with Folium

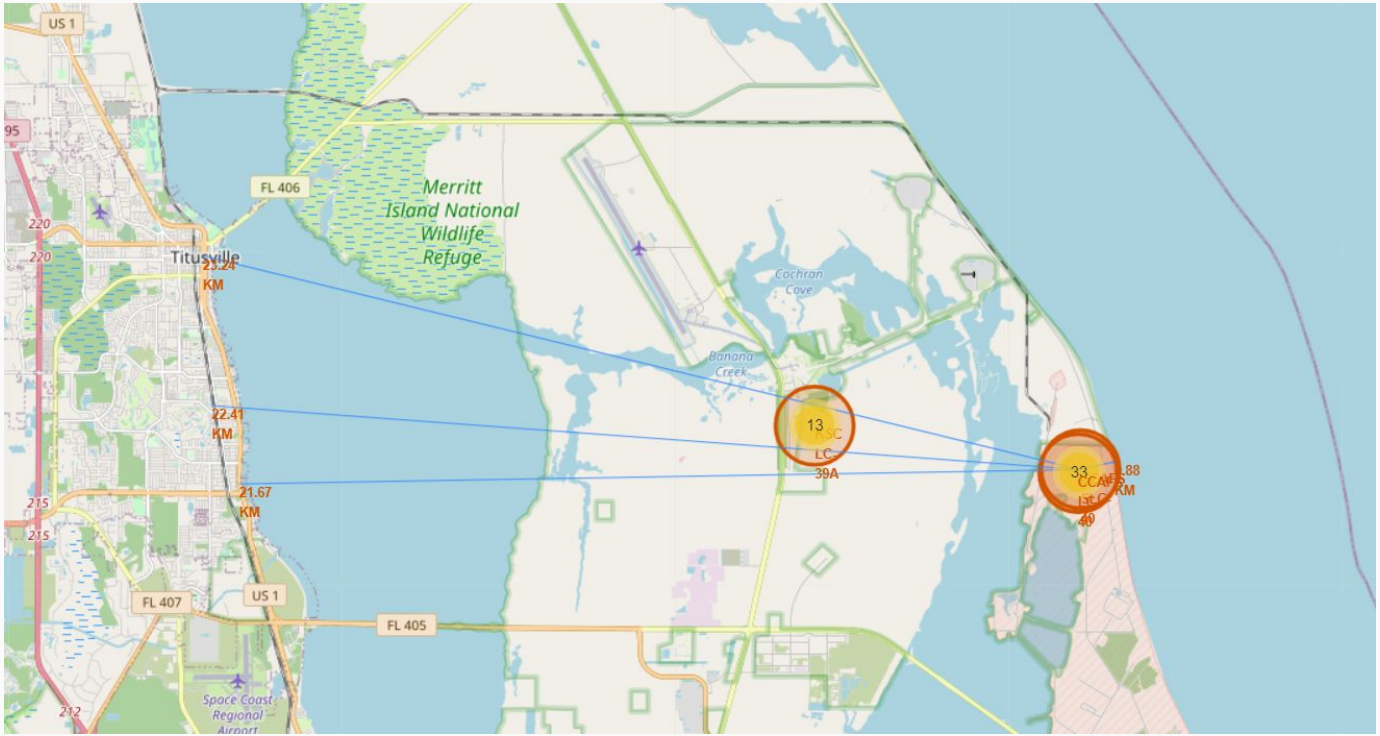
Next, let's try to enhance the map by adding the launch outcomes for each site, and see which sites have high success rates. Recall that data frame spacex_df has detailed launch records, and the `class` column indicates if this launch was successful or not.

	Launch Site	Lat	Long	class
46	KSC LC-39A	28.573255	-80.646895	1
47	KSC LC-39A	28.573255	-80.646895	1
48	KSC LC-39A	28.573255	-80.646895	1
49	CCAFS SLC-40	28.563197	-80.576820	1
50	CCAFS SLC-40	28.563197	-80.576820	1
51	CCAFS SLC-40	28.563197	-80.576820	0
52	CCAFS SLC-40	28.563197	-80.576820	0
53	CCAFS SLC-40	28.563197	-80.576820	0
54	CCAFS SLC-40	28.563197	-80.576820	1
55	CCAFS SLC-40	28.563197	-80.576820	0



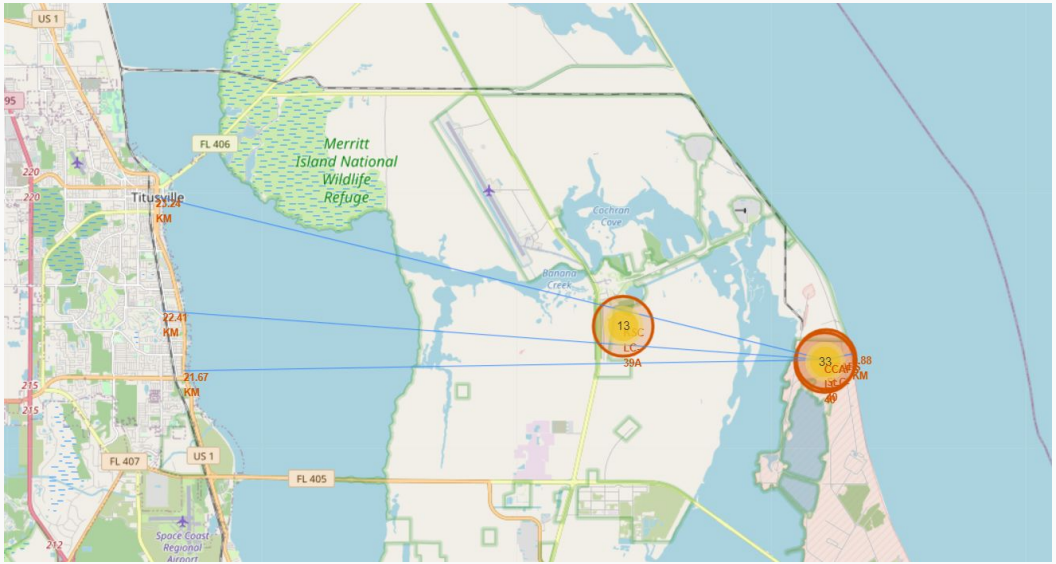
Analysis with Folium

Next, we need to explore and analyze the proximities of launch sites.



Analysis with Folium

Next, we need to explore and analyze the proximities of launch sites.
We can Observe that based on the data:
The launch sites are approximately closer to shores, which the loser it is the higher the success rate, also, the most launches comes from sites that are approximately closer to infrastructures, like rails, highways, airports, and they are relatively safely far from urban areas.



Machine Learning

Predictive Analysis Overview

- Objective: To predict the success of Falcon 9 first stage landings.
- Methodology: Utilized Logistic Regression, Support Vector Machine (SVM), Decision Trees, and K-Nearest Neighbors (KNN) for classification.

Feature Engineering and Data Splitting

- Features: Various parameters like flight number, date, payload mass, and orbit.
- Data Splitting: Used train-test split for model validation.

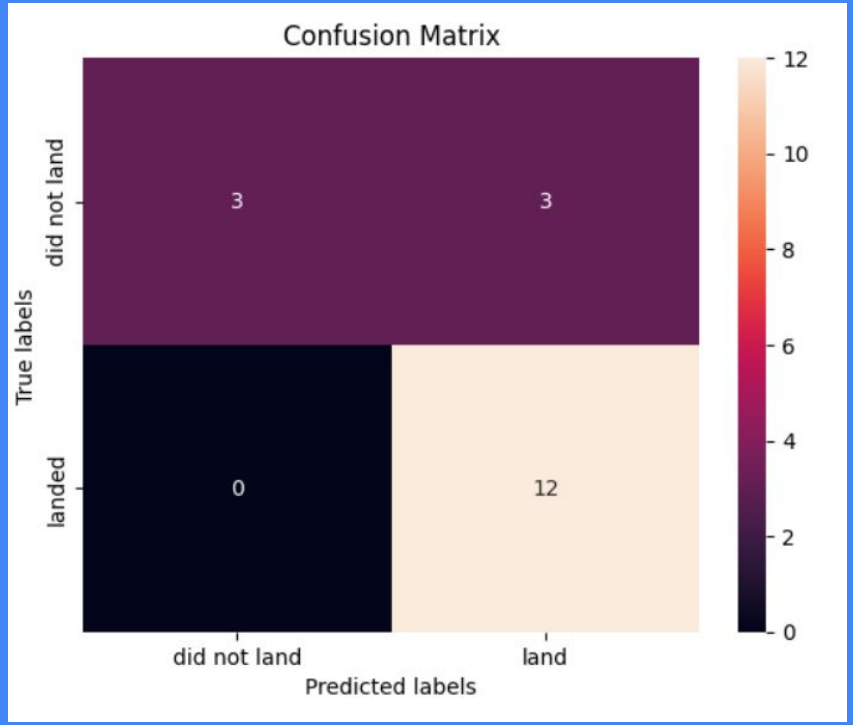
Machine Learning

Logistic Regression

Hyperparameter Tuning: Utilized GridSearchCV with cv = 10.

Best Parameters: C = 0.01, penalty = 'l2', solver = 'lbfgs'.

Test Accuracy: Approximately 0.8333.



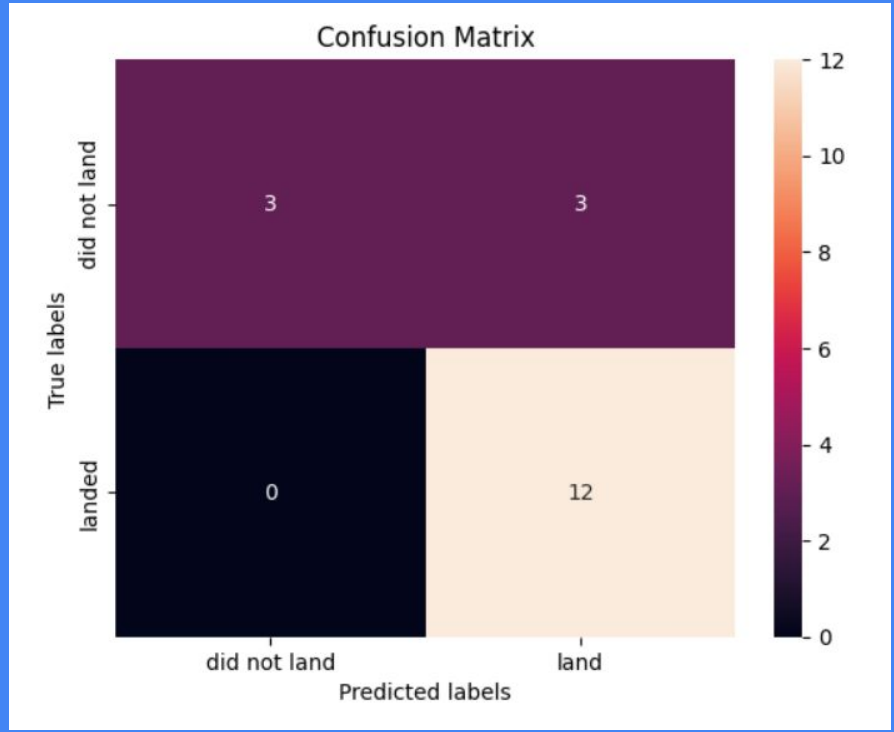
Machine Learning

Support Vector Machine (SVM)

Hyperparameter Tuning: Utilized GridSearchCV with cv = 10.

Best Parameters: C = 1.0, gamma = 0.0316, kernel = 'sigmoid'.

Test Accuracy: Approximately 0.8333.



Machine Learning

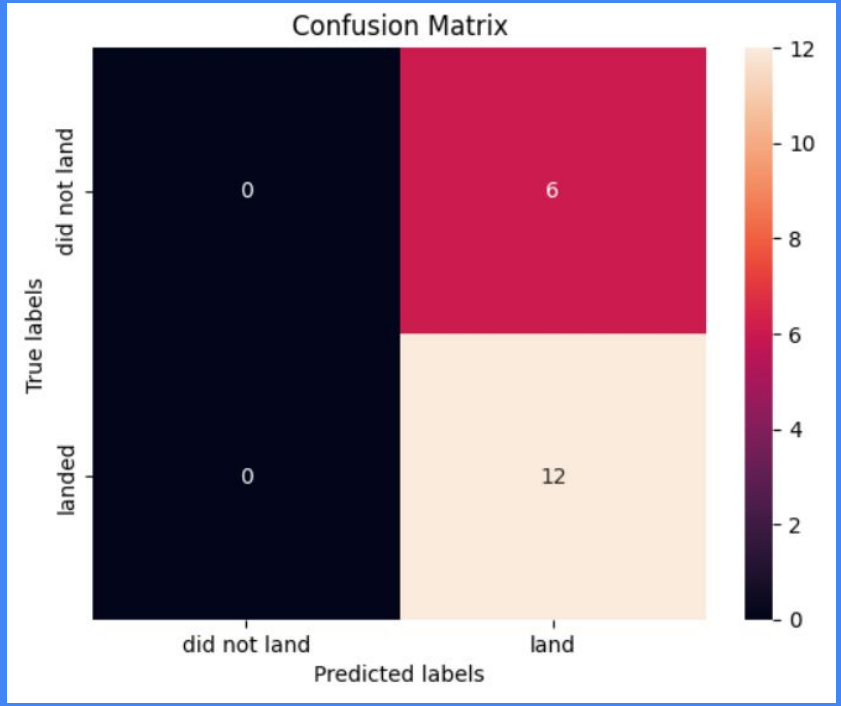
Decision Trees

Hyperparameter Tuning: Utilized GridSearchCV with cv = 10.

Best Parameters: Criteria = 'entropy', max depth = 2, etc.

Test Accuracy: Approximately 0.875.

Confusion Matrix: [Insert Confusion Matrix Plot]



Machine Learning

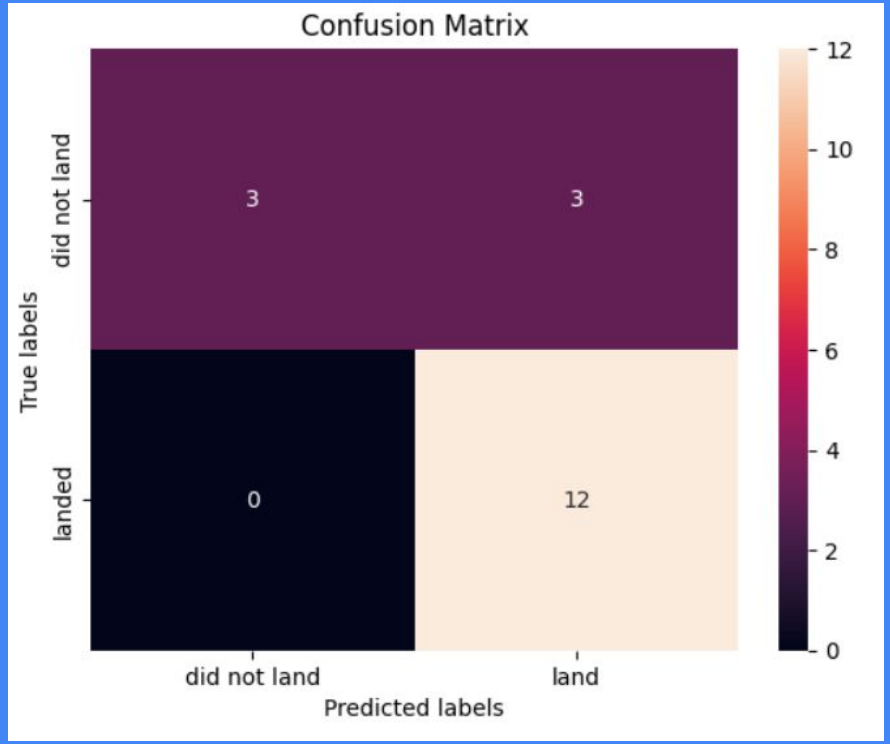
K-Nearest Neighbors (KNN)

Hyperparameter Tuning: Utilized GridSearchCV with cv = 10.

Best Parameters: Algorithm = 'auto', n_neighbors = 10, p = 1.

Test Accuracy: Approximately 0.8333.

Confusion Matrix: [Insert Confusion Matrix Plot]



Final Insights: Space X Falcon 9 First Stage Landing Prediction

Objectives:

The primary objective of this project was to predict whether the first stage of a SpaceX Falcon 9 rocket would successfully land after launch. This is crucial for SpaceX's business model, which relies heavily on reusability to lower launch costs.

Data and Preprocessing:

The dataset comprised variables like FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, and others. A total of 83 features were prepared and standardized, and the dataset was split into training and test sets.

Methods:

Four machine learning algorithms were employed:

- Logistic Regression
- Support Vector Machines (SVM)
- Decision Trees
- K-Nearest Neighbors (KNN)

The models were optimized using GridSearchCV for hyperparameter tuning and were evaluated based on their accuracy.

Final Insights: Space X Falcon 9 First Stage Landing Prediction

Key Findings:

Hyperparameter Tuning: The best parameters for each model were obtained. For example, the Logistic Regression model performed best with C=0.01, penalty='l2', and solver='lbfgs'.
Validation Accuracy: All models had a similar accuracy score during validation, ranging from 84.6% to 87.5%.
Test Accuracy: On the test set, three models (Logistic Regression, SVM, and KNN) had an accuracy of 83.3%, while Decision Trees lagged with a 66.7% accuracy.
Confusion Matrix: The major issue across all models was the number of false positives, indicating that the models may predict a successful landing when, in fact, it may not be successful.

Best Performing Model:

The Logistic Regression model had the highest test accuracy and is recommended for making predictions. However, it's worth noting that KNN and SVM also performed equally well on the test set.

Final Insights: Space X Falcon 9 First Stage Landing Prediction

Practical Implications:

The insights gained from this analysis have practical applications for not just SpaceX but also for other companies that may want to bid for rocket launches. Knowing the likelihood of a successful landing can help in risk assessment, mission planning, and cost estimation.

Future Work:

- Integrate more features like weather conditions, which could play a significant role in the success of the landing.
- Use ensemble methods to possibly improve accuracy.
- Address the issue of false positives through techniques like cost-sensitive learning or by assembling a more balanced dataset.

Conclusion:

Machine learning models can effectively predict the success of Falcon 9 first stage landings with an accuracy upwards of 83%. This not only affirms the capabilities of machine learning in this domain but also offers a data-driven approach to optimize rocket launches, thereby saving millions of dollars.

Thanks!

Hazem Haffouz