



Falcon 9 SpaceX Data Science Capstone Project

Analysis and Insights

KATA RAVI

10.08.2024



OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion



EXECUTIVE SUMMARY

- In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.

The main steps in this project include:

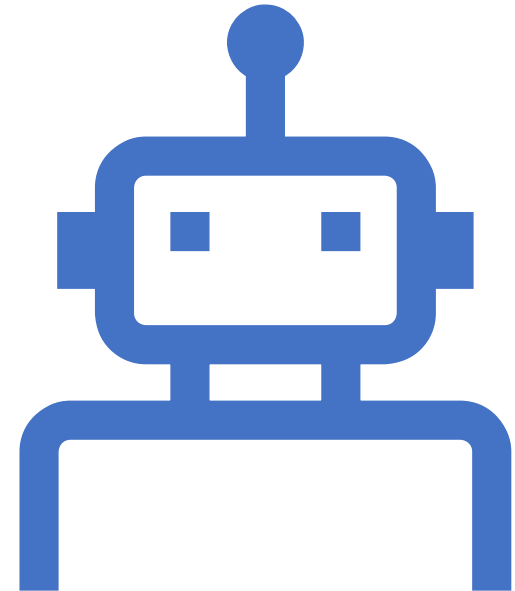
Data collection, wrangling, and formatting

Exploratory data analysis

Interactive data visualization

Machine learning prediction

- Our graphs show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure.
- It is also concluded that decision tree algorithm may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully



Introduction

The Falcon 9 is a partially reusable two-stage rocket designed and manufactured by SpaceX for the reliable and safe transport of satellites and the Dragon spacecraft into orbit. Since its debut in 2010, it has become known for its innovative reusable first stage, which can land back on Earth after launch to be refurbished and flown again. This capability has significantly reduced the cost of access to space. Falcon 9 has supported numerous missions, including commercial satellite deployments, cargo resupply missions to the International Space Station, and crewed spaceflights under NASA's Commercial Crew Program



METHODOLOGY

The overall methodology includes:

1. Data collection, wrangling, and formatting, using:

- SpaceX API
- Web scraping

2. Exploratory data analysis (EDA), using:

- Pandas and NumPy
- SQL

3. Data visualization, using:

- Matplotlib and Seaborn
- Folium
- Dash

4. Machine learning prediction, using

- Logistic regression
- Support vector machine (SVM)

- Decision tree

- K-nearest neighbours (KNN)

METHODOLOGY

Data collection, wrangling, and formatting

- SpaceX API
- The API used is <https://api.spacexdata.com/v4/rockets/>.
- The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
- Every missing value in the data is replaced the mean the column that the missing value belongs to.
- We end up with 90 rows or instances and 17 columns or features. The picture below shows the first few rows of the data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs		LandingPad	Block	Reuse
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False		None	1.0	
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False		None	1.0	
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False		None	1.0	
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False		None	1.0	
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False		None	1.0	

METHODOLOGY

Data collection, wrangling, and formatting

Data Wrangling

```
28]: data_falcon9.isnull().sum()
```

```
28]: FlightNumber      0
      Date              0
      BoosterVersion    0
      PayloadMass       5
      Orbit             0
      LaunchSite        0
      Outcome           0
      Flights           0
      GridFins          0
      Reused            0
      Legs              0
      LandingPad        26
      Block             0
      ReusedCount       0
      Serial            0
      Longitude         0
      Latitude          0
      dtype: int64
```

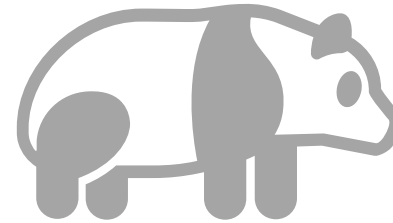
```
[29]: FlightNumber      0
      Date              0
      BoosterVersion    0
      PayloadMass       0
      Orbit             0
      LaunchSite        0
      Outcome           0
      Flights           0
      GridFins          0
      Reused            0
      Legs              0
      LandingPad        26
      Block             0
      ReusedCount       0
      Serial            0
      Longitude         0
      Latitude          0
      dtype: int64
```

METHODOLOGY

Web Scrapping from Wikipedia:



BeautifulSoup was used to scrape Falcon 9 launch records from Wikipedia.



The HTML table containing the launch records was parsed and converted into a pandas DataFrame.

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10
5	6	VAFB	CASSIOPE	500 kg	Polar orbit	MDA	Success	F9 v1.1B1003	Uncontrolled	29 September 2013	16:00
6	7	CCAFS	SES-8	3,170 kg	GTO	SES	Success	F9 v1.1	No attempt	3 December 2013	22:41
7	8	CCAFS	Thaicom 6	3,325 kg	GTO	Thaicom	Success	F9 v1.1	No attempt	6 January 2014	22:06
8	9	Cape Canaveral	SpaceX CRS-3	2,296 kg	LEO	NASA (CRS)	Success\n	F9 v1.1	Controlled	18 April 2014	19:25

METHODOLOGY

Web Scrapping from Wikipedia:



METHODOLOGY

Data collection, wrangling, and formatting

- The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.
- An extra column called 'Class' is also added to the data frame. The column 'Class' contains 0 if a given launch is failed and 1 if it is successful.
- In the end, we end up with 90 rows or instances and 83 columns or features.

METHODOLOGY

Exploratory data analysis (EDA)

- **Visualization:**
- Used various plots to visualize the relationship between different features like flight number, launch site, payload mass, and success rate.
- Plotted success rate trends over the years and success rates of different orbit types

METHODOLOGY

Machine Learning Prediction

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase include the following steps:
 - Standardizing the data
 - Splitting the data into training and test data
 - Creating machine learning models, which include:
 - Logistic regression
 - Support vector machine (SVM)
 - Decision tree
 - K nearest neighbors (KNN)
 - Fit the models on the training set
 - Find the best combination of hyperparameters for each model
 - Evaluate the models based on their accuracy scores and confusion matrix

RESULTS

- The results are split into 5 sections:
- SQL (EDA with SQL)
- Matplotlib and Seaborn (EDA with Visualization)
- Folium
- Dash
- Predictive Analysis
- In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

RESULTS

EDA with SQL

The names of the unique launch sites in the space mission

Done.

: **Launch_Site**

CCAFS LC-40

VAFB SLC-4E

KSC LC-39A

CCAFS SLC-40

5 records where launch sites begin with 'CCA'

Done.

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

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

total_payload_mass

45596

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

average payload mass carried by booster version F9 v1.1

2534

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

2010-06-04

RESULTS

EDA with SQL

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

Booster_Version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Done.

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

RESULTS

EDA with SQL

The names of the booster versions which have carried the maximum payload mass

: **Booster_Version**

F9 B5 B1048.4

F9 B5 B1049.4

F9 B5 B1051.3

F9 B5 B1056.4

F9 B5 B1049.5

RESULTS

EDA with SQL

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

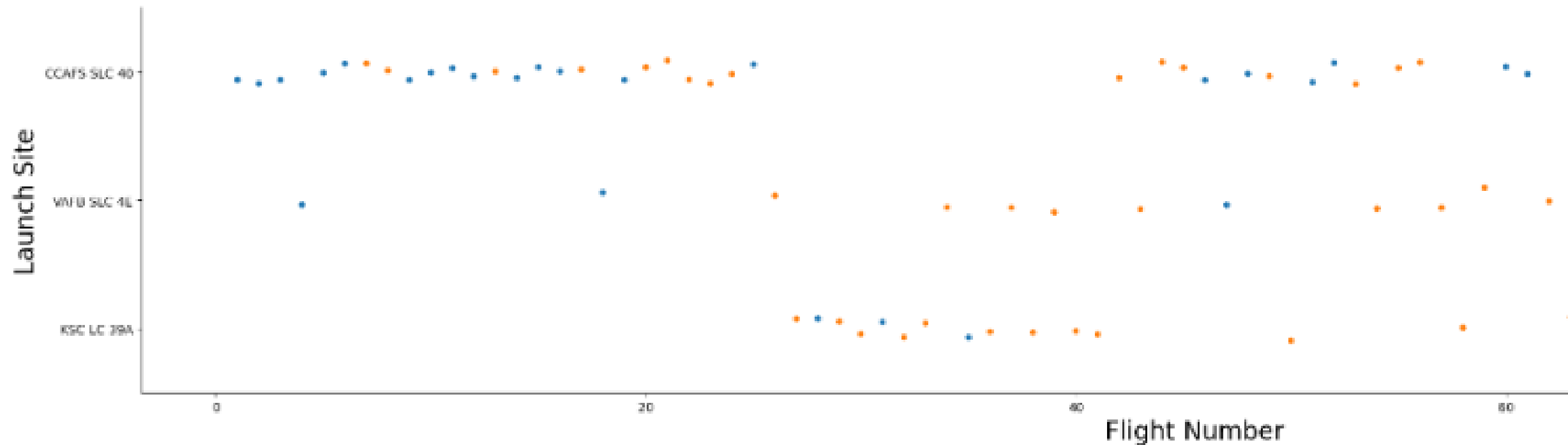
The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Done .

Landing_Outcome	count
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

Done .

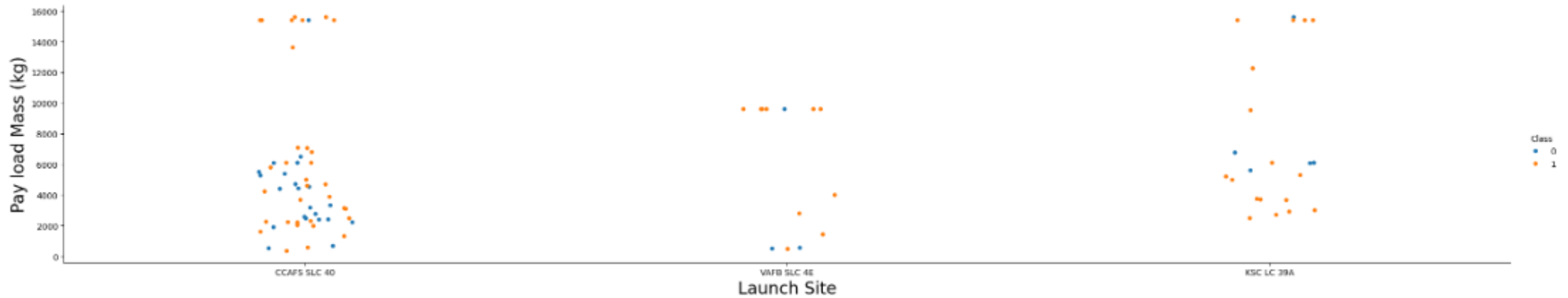
month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40



- The relationship between flight number and launch site

RESULTS

Matplotlib and Seaborn (EDA with Visualization)



RESULTS

Matplotlib and Seaborn (EDA with Visualization)

The relationship between payload mass and launch site

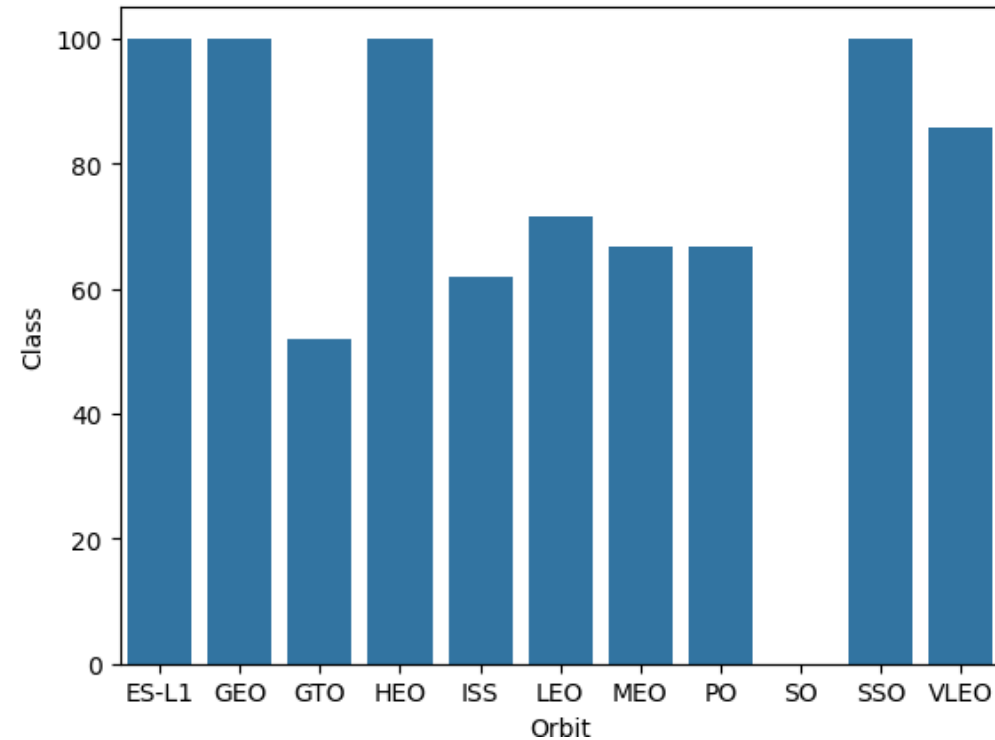


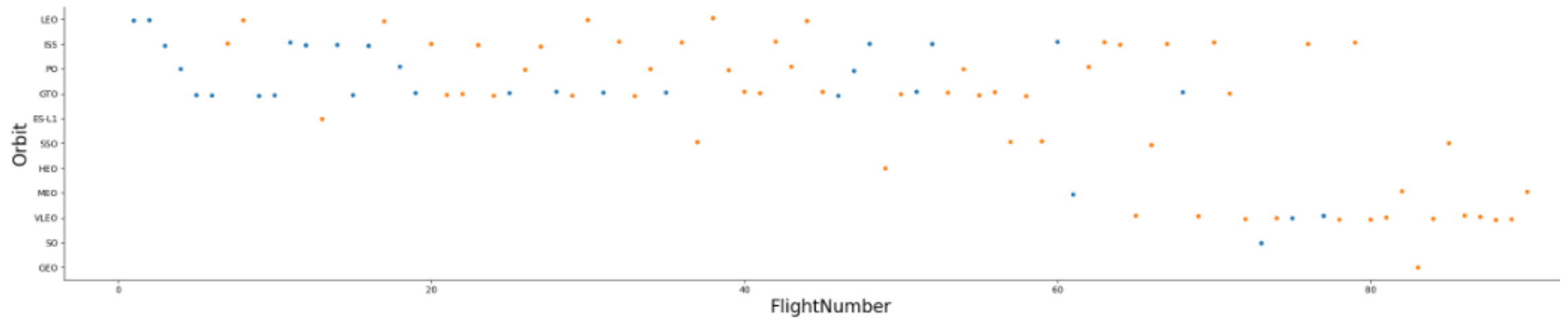
RESULTS

Matplotlib and Seaborn (EDA with Visualization)

The relationship between success rate and orbit type

<Axes: xlabel='Orbit', ylabel='Class'>

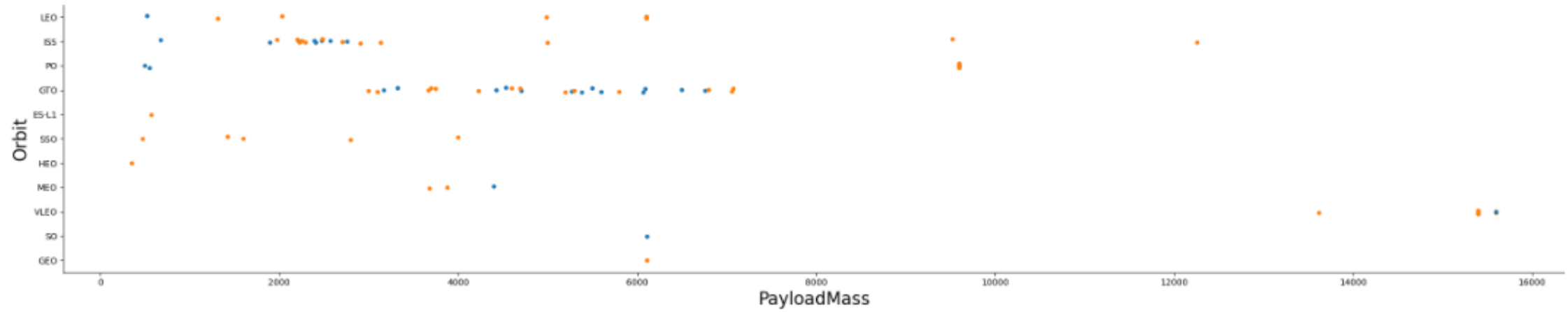




RESULTS

Matplotlib and Seaborn (EDA with Visualization)

The relationship between flight number and orbit type



RESULTS

Matplotlib and Seaborn (EDA with Visualization)

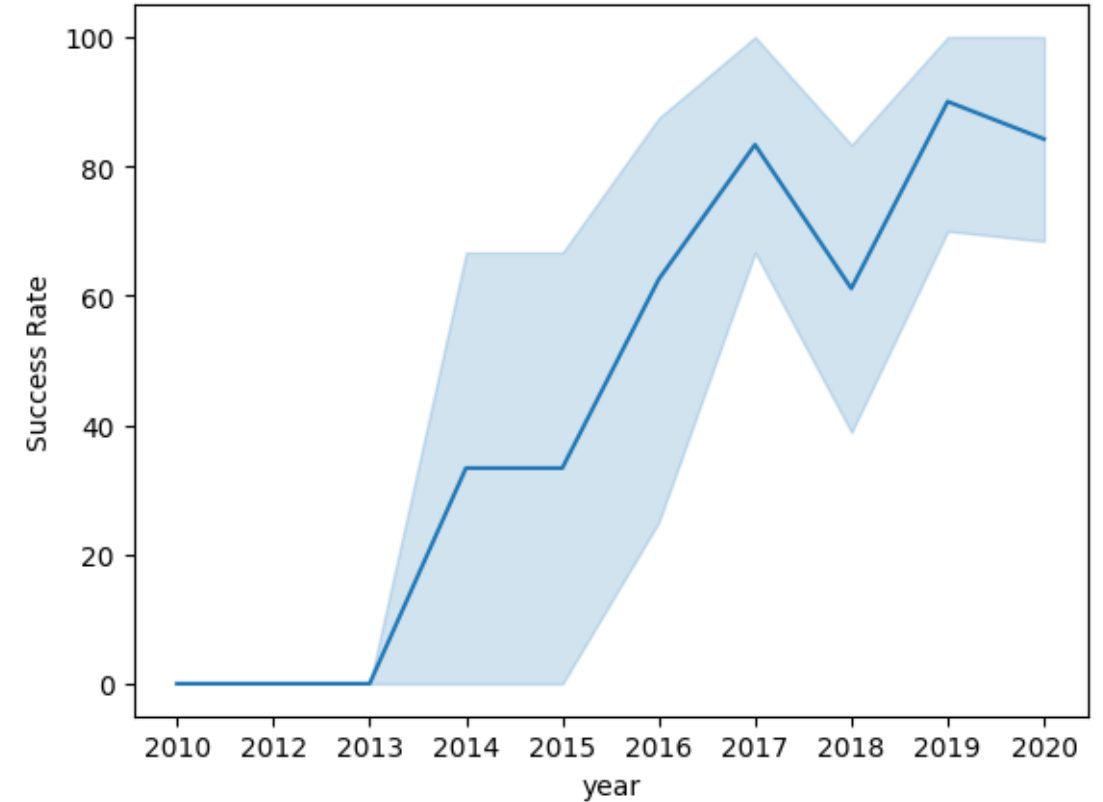
The relationship between payload mass and orbit type



RESULTS

Matplotlib and Seaborn (EDA with Visualization)

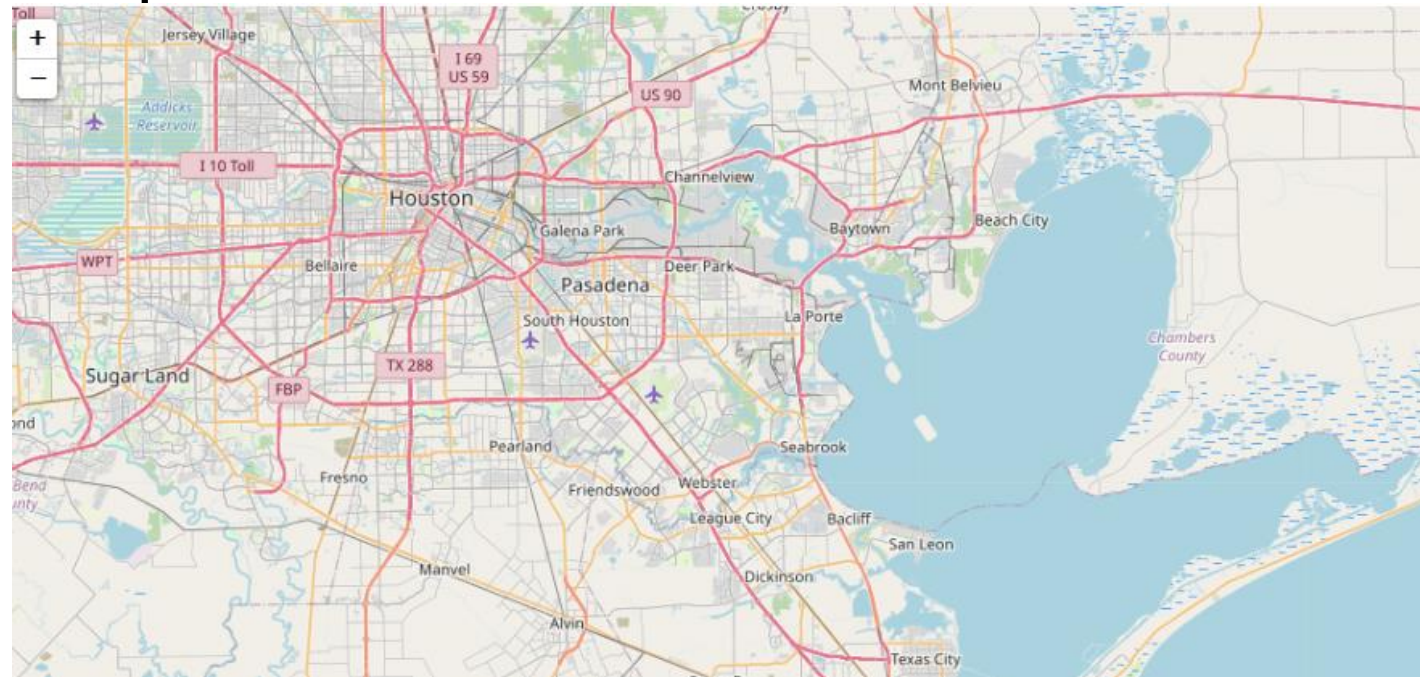
The launch success yearly trend



RESULTS

Folium

All launch sites on map

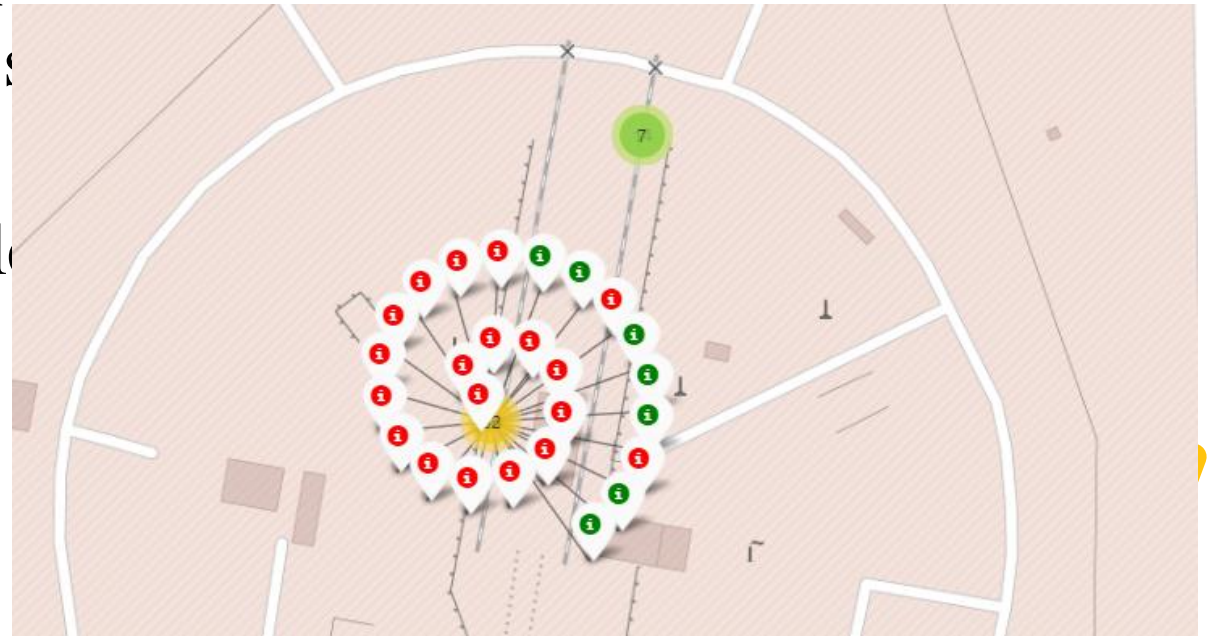


RESULTS

Folium

If we zoom in on one of the launch site, we can see green and red tags

Each green tag represents a successful launch while each red tag represents a failed launch



RESULTS

Folium

The picture below shows the distance between the CCAFS SLC-40 launch site and the nearest Coastline

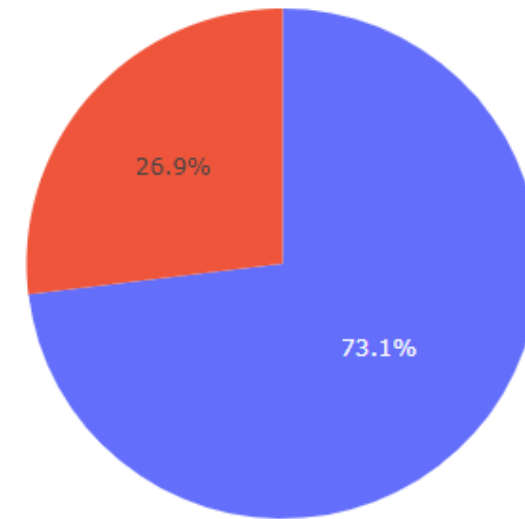


RESULTS

Dash

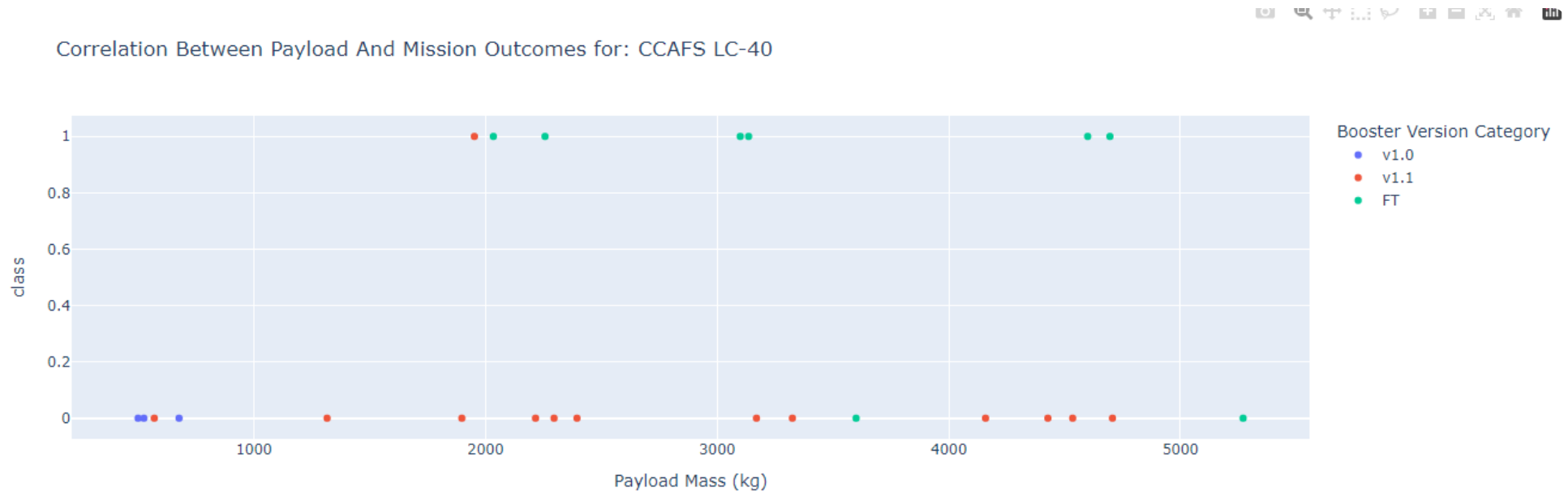
- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
 - 0 represents failed launches while 1 represents successful launches.
- We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.

X Launch Records D



RESULTS

Dash



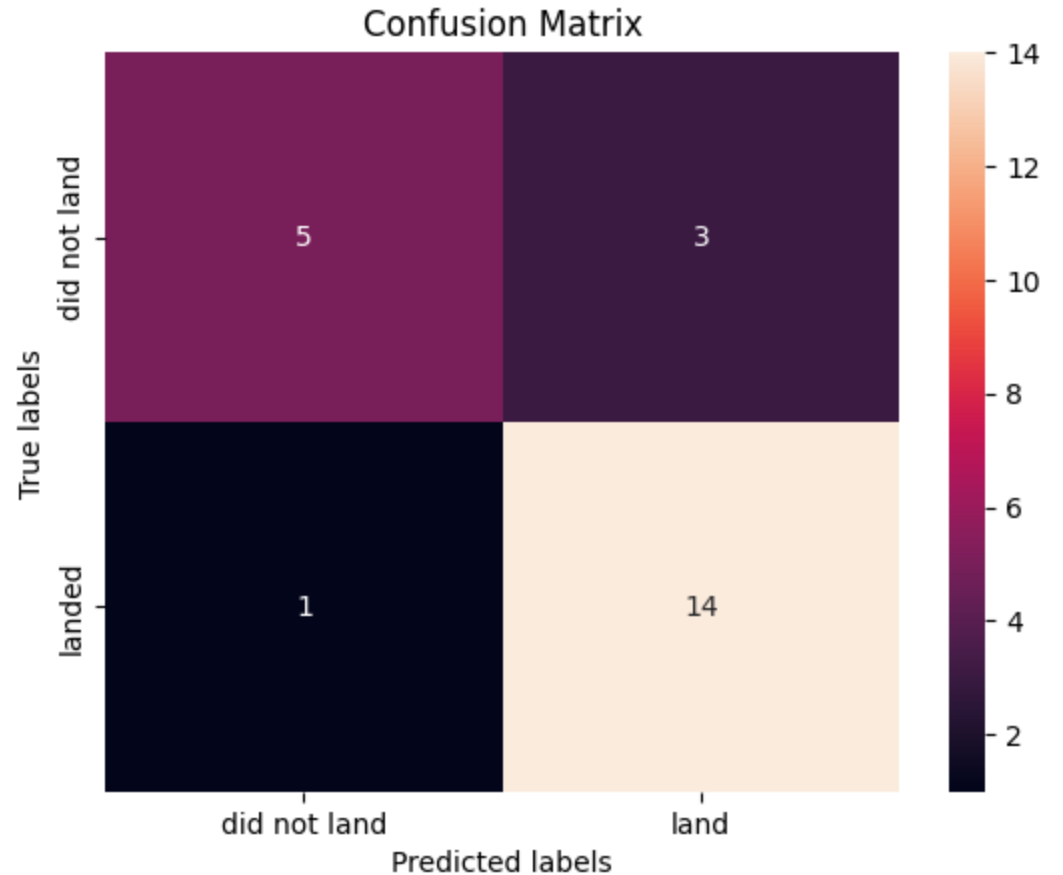
- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.

RESULTS

Predictive Analysis

Logistic regression

- GridSearchCV best score:
0.8238095238095238
- Accuracy score on test set:
0.8260869565217391
- Confusion matrix



RESULTS

Predictive Analysis

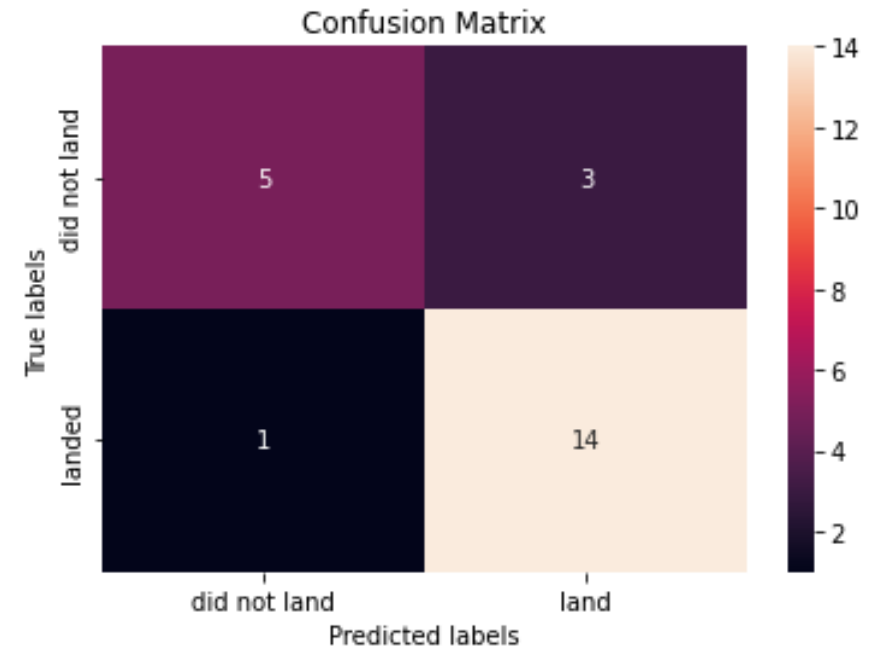
Support vector machine (SVM)

- GridSearchCV best score:
0.838095238095238

- Accuracy score on test set:

test set accuracy :
0.8260869565217391

- Confusion matrix:

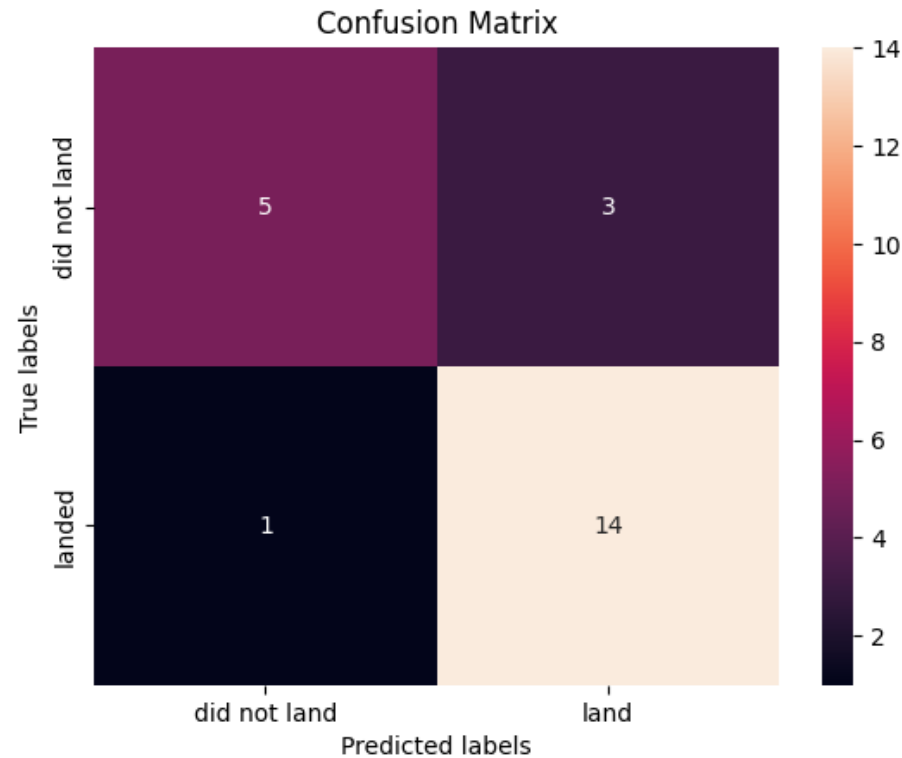


RESULTS

Predictive Analysis

Decision tree

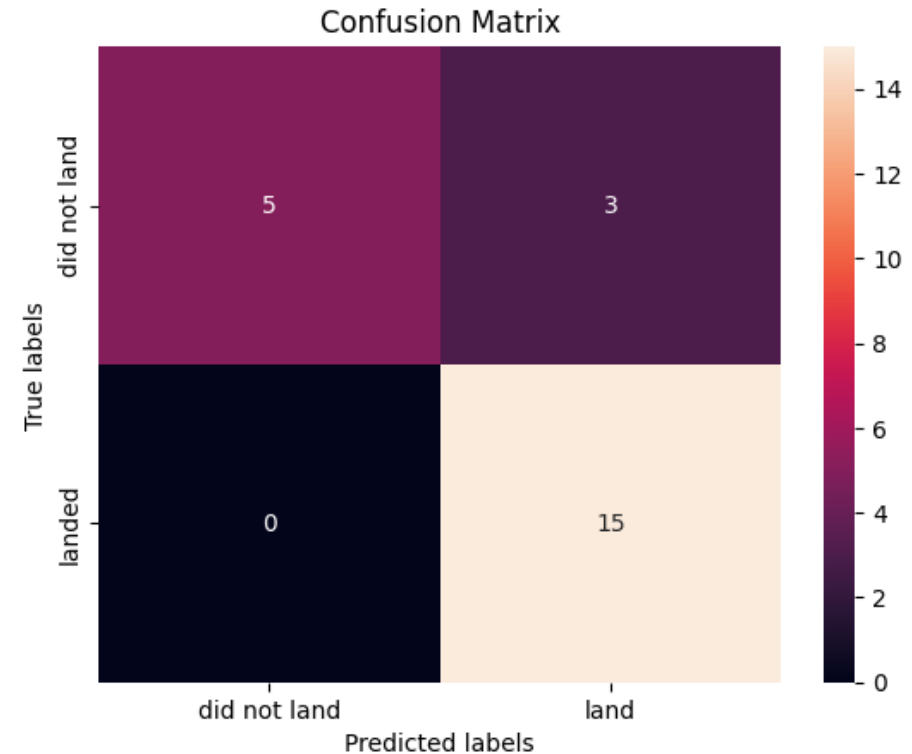
- GridSearchCV best score:
0.8833333333333332
- Accuracy score on test set:
0.8260869565217391
- Confusion matrix



RESULTS

Predictive Analysis

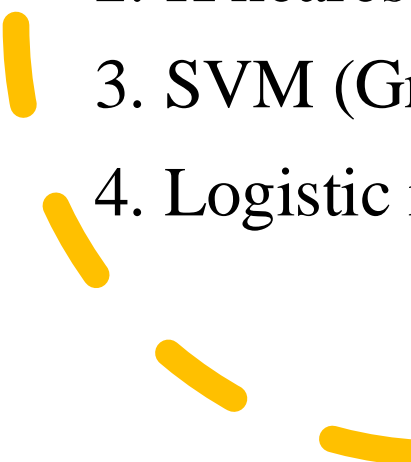
- K nearest neighbors (KNN)
- GridSearchCV best score:
0.85
- Accuracy score on test set:
0.8695652173913043
- Confusion matrix:





RESULTS

Predictive Analysis

- Putting the results of all 4 models: Best Model Decision Tree
1. Decision tree (GridSearchCV best score: 0.8833333333333333332)
 2. K nearest neighbors, KNN (GridSearchCV best score:0.85)
 3. SVM (GridSearchCV best score: 0.838095238095238)
 4. Logistic regression (GridSearchCV best score: 0.8238095238095238)
- 



Machine Learning

We must also look at a high volume of connected data (from both internal and external sources) to gain a more complete picture of future outcomes. Every area of a business can derive value from predictive modeling. Using robust algorithms

CONCLUSION

The project demonstrated that various factors like flight number, payload mass, and orbit type significantly influence the success of Falcon 9's first stage landing. The success rate of launches has been improving over the years, and certain orbits have higher success rates. The Decision Tree Classifier was the most effective model for predicting landing success. These insights can help other companies to strategize and compete with SpaceX

innovative insights

Exploratory Data Analysis:

- Found that higher flight numbers at a launch site correlate with higher success rates.
- Certain orbits like ES-L1, GEO, HEO, SSO, and VLEO have higher success rates.
- Success rate has been increasing since 2013.

Interactive Analytics:

- Mapped all launch sites and their proximity to landmarks such as railways and highways.
- Visualized success rates of different launch sites with color-coded markers on a map.

Predictive Analysis:

- Decision Tree Classifier emerged as the best model for predicting landing outcomes.
- Confusion matrix analysis showed high accuracy, though there were some false positives.