IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

César CEBALLOS-CASTANO
18/01/2024

Outline

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

Executive Summary

In this capstone project, the main objective is to predict the successful landing of the SpaceX Falcon 9 first stage, with the ultimate goal of determining the cost of a launch. The approach involves leveraging various machine learning classification algorithms. The project follows a structured methodology, including phases such as Data Collection, Data Wrangling and Preprocessing, Exploratory Data Analysis, Data Visualization, and ultimately, Machine Learning Prediction.

Throughout the investigation, the analysis suggests that certain features related to rocket launches exhibit a correlation with the success or failure of these launches. The project employs different classification algorithms to make predictions. In the conclusion, it is highlighted that the Decision Tree algorithm appears to be the most suitable for addressing the specific problem of predicting the successful landing of the Falcon 9 first stage.

Introduction

In this capstone project, our primary objective is to forecast the successful landing of the Falcon 9 first stage. SpaceX takes pride in its groundbreaking approach of reusing the first stage of a rocket launch, a cost-saving strategy that sets them apart. The company prominently advertises on its website that its rocket launches cost a mere 62 million dollars, a stark contrast to competitors whose prices soar upwards of 165 million dollars. The crux of these savings lies in the remarkable reusability of the first stage.

The ability to predict whether the first stage will successfully land holds pivotal significance, as it directly influences the overall cost of a launch. This valuable information becomes a strategic tool, especially for companies considering bidding against SpaceX for a rocket launch contract. The central question guiding our investigation revolves around a set of features related to a Falcon 9 rocket launch: Will the first stage of the rocket land successfully? Unraveling the intricacies of this query through machine learning and data analysis stands at the forefront of our capstone project.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - How data was collected
- Perform data wrangling
  - How data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- API : spacex_url=« https://api.spacexdata.com/v4/launches/past »
  - Required the data from Space API
  - Clean the data
- Web Page : https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
  - Extract a Falcon 9 launch records HTML table from Wikipedia (using BeautifulSoup)
  - Parse the table and convert it into a Pandas data frame

Steps :

1. Request and parse the SpaceX launch data using the GET request
2. Normalize JSON response into a dataframe
3. Filter dataframe to only Falcon 9` launches and data wrangling
4. Handle missing values
5. Export to csv

Data Collection – SpaceX API

1 Getting Response from API

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

2 Converting Response to a .json file

```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

3 Apply custom functions to clean data

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

```
getBoosterVersion(data)
```

4 Assign list to dictionary then dataframe

```
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}
```

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

5 Filter dataframe and export to flat file (.csv)

```
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
```

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

Data Collection

Data Collection - Scraping

1 Getting Response from HTML

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

2 Creating BeautifulSoup Object

```
soup = BeautifulSoup(page.text, 'html.parser')
```

3 Finding tables

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

4 Getting column names

```
column_names = []
temp = soup.find_all('th')
for x in range(len(temp)):
    try:
     name = extract_column_from_header(temp[x])
     if (name is not None and len(name) > 0):
        column_names.append(name)
    except:
     pass
```

5 Creation of dictionary

```
launch_dict= dict.fromkeys(column_names)

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


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'] = []
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

6 Appending data to keys

```
extracted_row = 0
#Extract each table
for table_number,table in enumerate(
   # get table row
    for rows in table.find_all("tr")
        #check to see if first table
```

7 Converting dictionary to dataframe

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

8 Dataframe to .CSV

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

[Data Collection Scraping]()

Data Wrangling

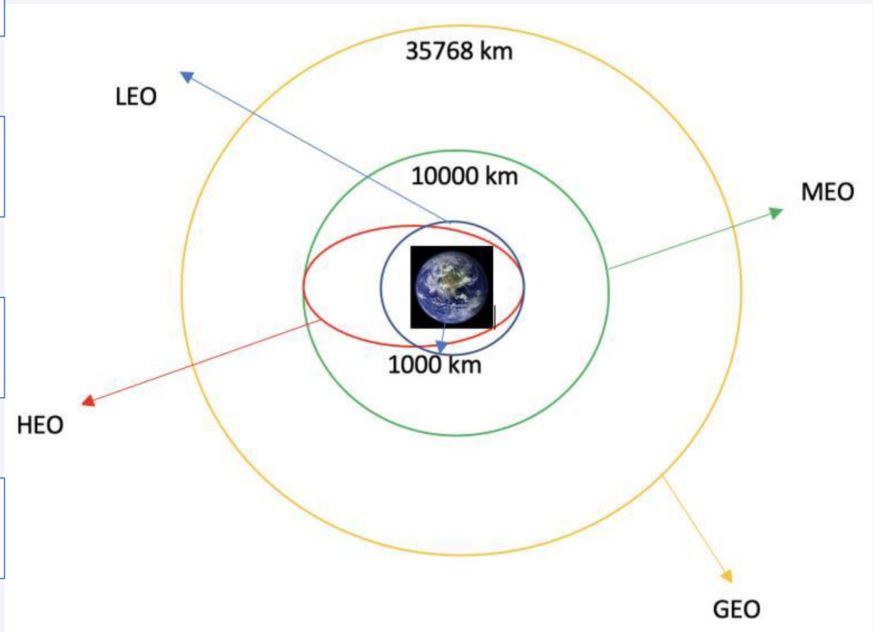
1. Calculate the number of launches on each site

2. Calculate the number and occurrence of each orbit

3. Calculate the number and occurence of mission outcome per orbit type

4. Create a landing outcome label from Outcome column using one-hot encoding

5. Export to CSV



[Data Wrangling](#)

EDA with Data Visualization

The relationship between :

Scatter Plot

- Flight Number and Launch Site
- Flight Number and Orbit type
- Payload and Orbit type
- Payload and Launch Site

Scatter plots are employed to depict the correlation between two variables, essentially showcasing how changes in one variable correspond to changes in another. Scatter plots allow for a comprehensive representation of the data set and aiding in the identification of patterns or trends in the relationship between the variables being analyzed.

Bar Plot

- Success rate of each orbit type

Bar diagrams are powerful tools for visually comparing data across diverse groups, utilizing categories and discrete values on respective axes to depict relationships. Primarily designed for quick comprehension of group comparisons, these versatile bar charts can effectively illustrate significant data changes over time, providing valuable insights into relationships and trends within datasets.

Line Chart

- The launch success yearly trend

Line graphs are valuable for presenting data variables and trends with clarity, enabling predictions about outcomes not yet recorded based on the observed patterns.

[Data Visualization]

EDA with SQL

- Displaying the names of the unique launch sites in the space mission
- Displaying 5 records where launch sites begin with the string 'KSC'
- Displaying the total payload mass carried by boosters launched by NASA (CRS)
- Displaying average payload mass carried by booster version F9 v1.1
- Listing the date where the successful landing outcome in drone ship was achieved.
- Listing the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass.
- Listing the records which will display the month names, successful landing_outcomesin ground pad ,booster versions, launch_sitefor the months in year 2017
- Ranking the count of successful landing_outcomesbetween the date 2010-06-04 and 2017-03-20 in descending order.

SQL

Build an Interactive Map with Folium

We made stuff and put it on a Folium map. Markers were used to show where all the launch sites are, and whether they were successful or not. Lines were drawn to figure out how far each launch site is from its nearby spots.

We assigned the dataframe launch_outcomes (failures, successes) to classes 0 and 1 with Green and Red markers on the map in a MarkerCluster()

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities
    - Whether it is close to the coast
    - Whether it is close to the railway
    - Whether it is close to the highway
    - Whether it is close to the city

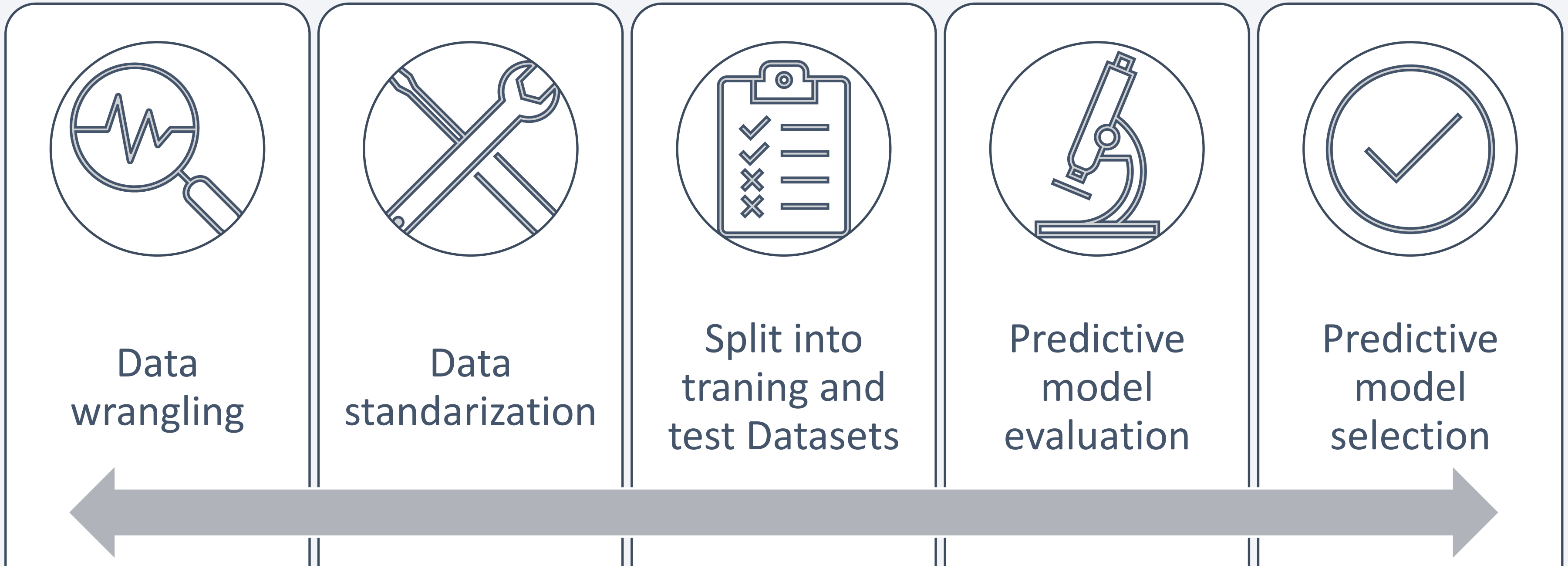
Interactive Map

Build a Dashboard with Plotly Dash

- A Scatter Graph illustrates the correlation between Outcome and Payload Mass (Kg) for various Booster Versions. This method excels at showcasing non-linear patterns, providing a clear view of data range from minimum to maximum values. It ensures straightforward observation and reading of the relationship between the two variables.

  - A launch site drop-down input component
  - A success-pie-chart based on the selected site dropdown
  - A range slicer to select payload
  - A success-payload-scatter-chart scatter plot based on the selected site dropdown

Predictive Analysis (Classification)



Load our dataset into NumPy and Pandas
Transform Data
Split our data into training and test data sets
Check how many test samples we have
Decide which type of machine learning algorithms we want to use
Set our parameters and algorithms to GridSearchCV
Fit our datasets into the GridSearchCVobjects and train our dataset.

Logistic regression
Support vector machine
Decision tree classifer
K-nearest neighnors

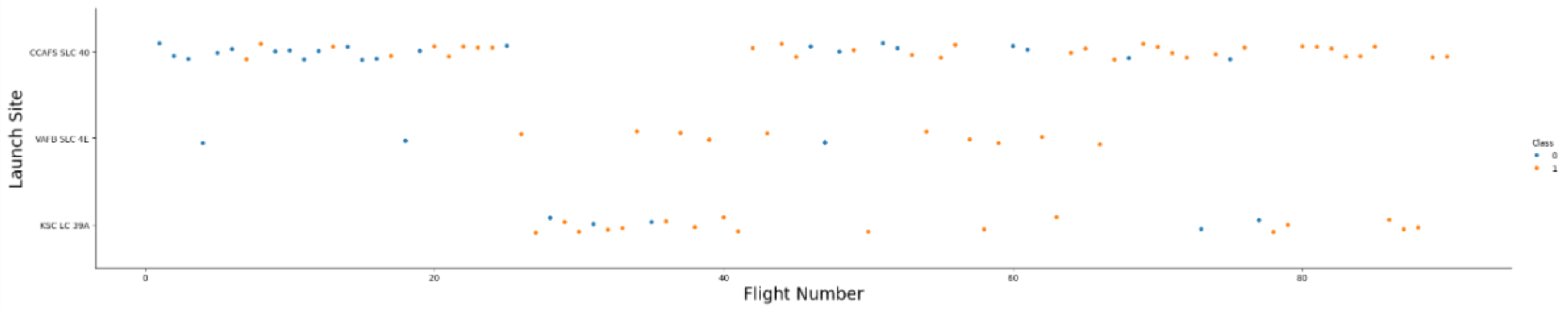
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

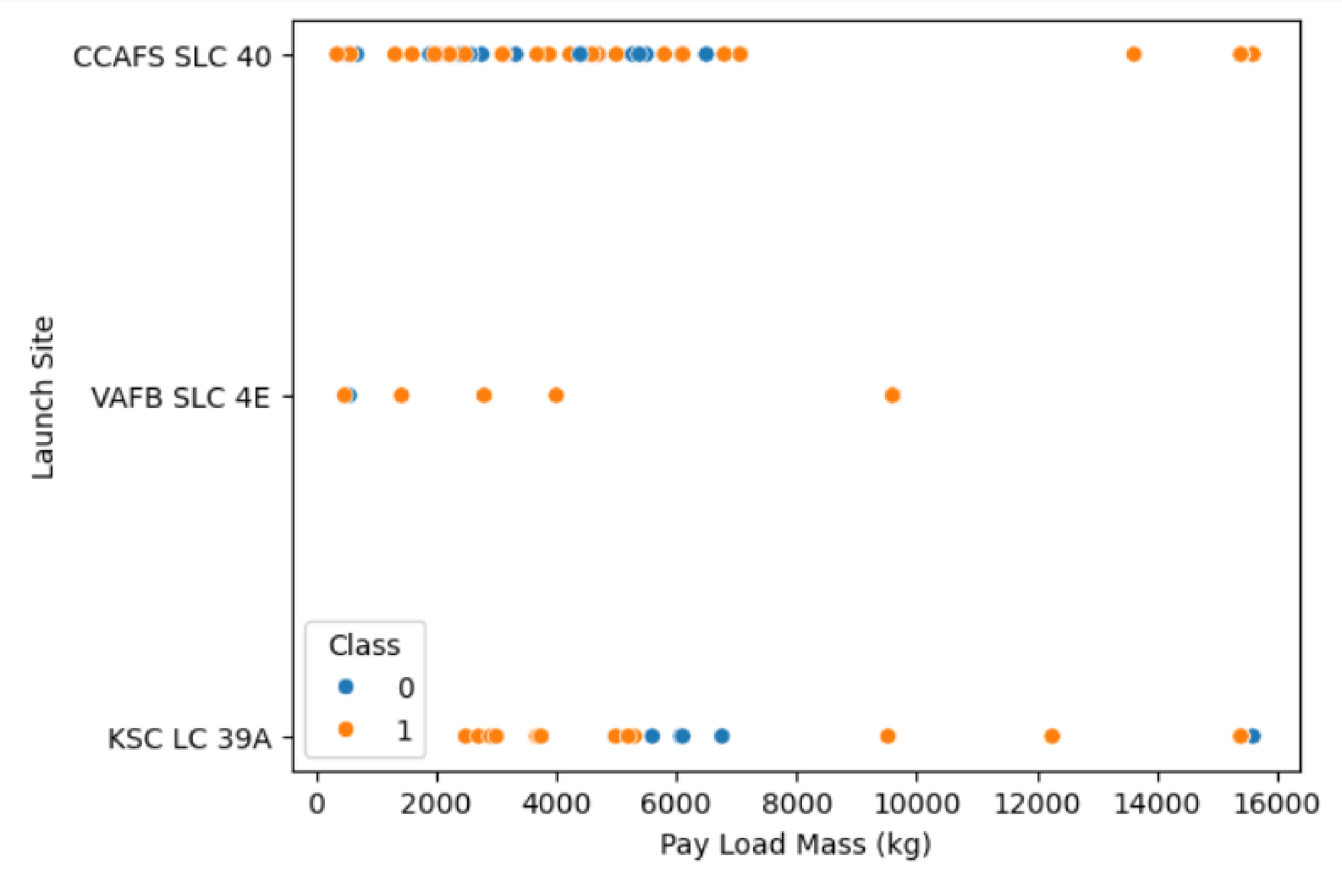


This figure shows that the success rate increased as the number of flights increased.

The blue dots represent the successful launches while the red dot represent unsuccessful luanches.

There seems to be an increase in successful flights after the 40th launch.

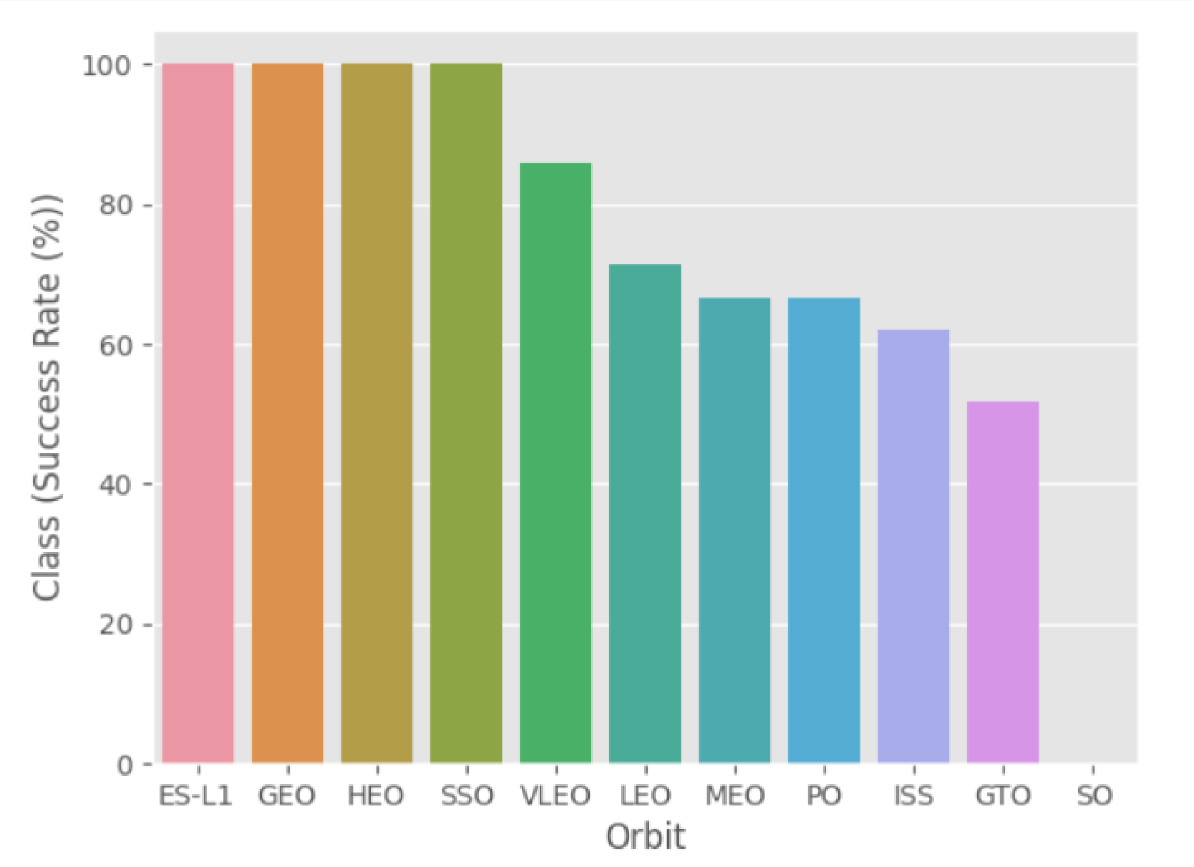
Payload vs. Launch Site



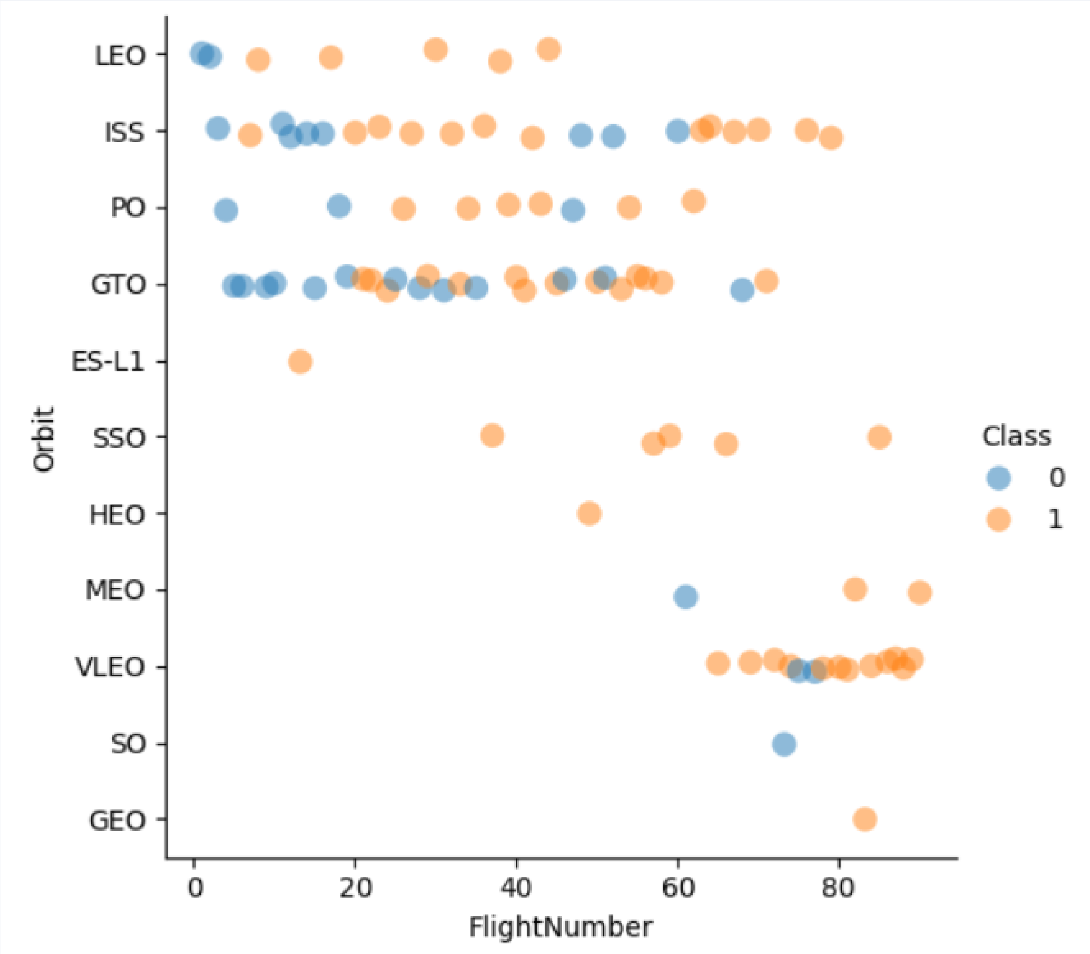
A higher payload mass at Launch Site CCAFS SLC 40 correlates with an increased success rate for the rocket. However, the visualization doesn't distinctly reveal a pattern to determine if the launch site is significantly dependent on payload mass for a successful launch decision.

Success Rate vs. Orbit Type

Orbits ES-L1, GEO, HEO & SSO have the highest success rates at 100%, with SO orbit having the lowest success rate at ~50%. Orbit SO has 0% success rate.



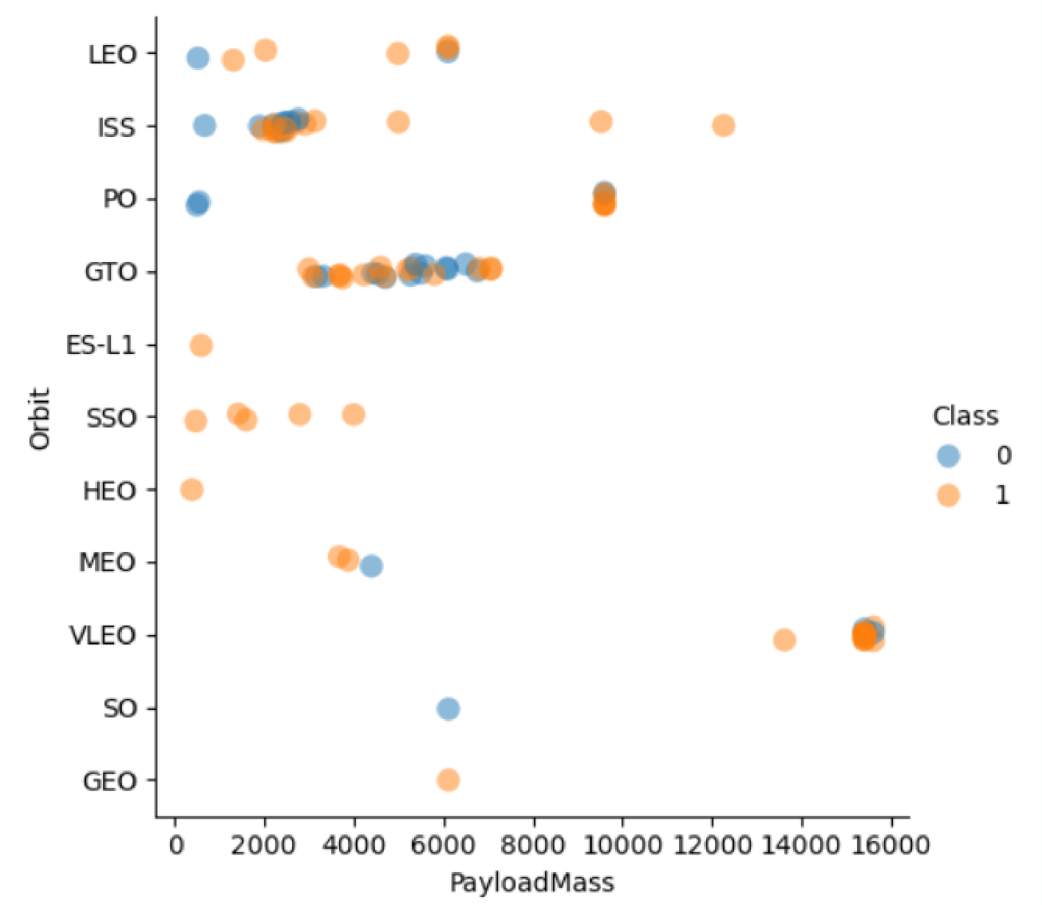
Flight Number vs. Orbit Type



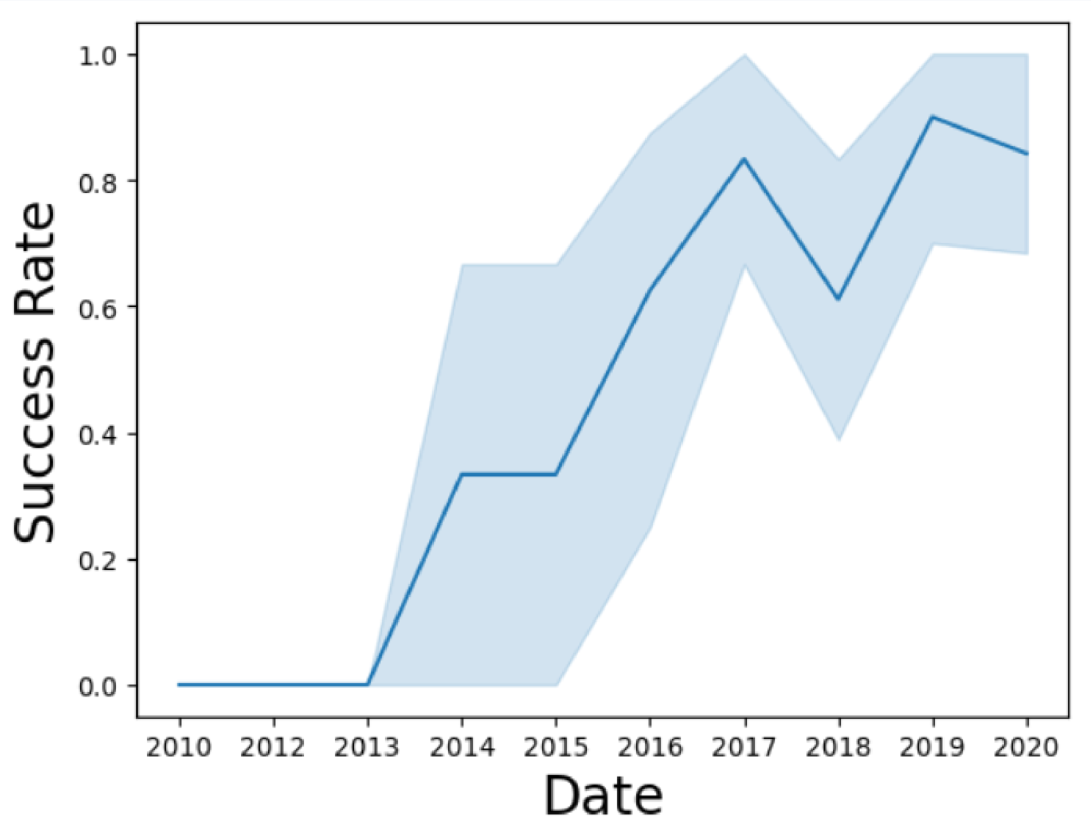
In the LEO orbit, success seems linked to the number of flights, while in the GTO orbit, there appears to be no discernible relationship between success and flight number.

Payload vs. Orbit Type

Note that heavy payloads negatively impact GTO orbits but have a positive effect on GTO and Polar LEO (ISS) orbits.



Launch Success Yearly Trend



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

All Launch Site Names

SQL query : **select Distinct(LAUNCH_SITE) from tblSpaceX**

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

Using the word DISTINCT in the query means that it will only show Unique values in the Launch_Site column from tblSpaceX

Launch Site Names Begin with 'CCA'

SQL query : **SELECT * from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5**

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

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

Time JTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
5:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
3:00	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)
4:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
5:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
0:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Using the query, it will only show 5 records from tblSpaceX and LIKE keyword has a wild card with the words the percentage in the end suggests that the Launch_Site name must start with CCA.

Total Payload Mass

SQL query : %sql select sum(PAYLOAD_MASS__KG_) as TotalPayloadMass from SPACEXTBL where Customer =='NASA (CRS)’

TotalPayloadMass
45596

Using the function ***SUM*** summates the total in the column ***PAYLOAD_MASS_KG_***. The ***WHERE*** clause filters the dataset to only perform calculations on ***Customer NASA (CRS)***

Average Payload Mass by F9 v1.1

SQL query : %sql select avg(PAYLOAD_MASS__KG_) as Avg_payload from SPACEXTBL where Booster_version Like 'F9 v1.1’;

Avg_payload
2928.4

Using the function AVG works out the average in the column PAYLOAD_MASS_KG_

The WHERE clause filters the dataset to only perform calculations on Booster_version F9 v1.1

First Successful Ground Landing Date

SQL query : select min(DATE) from SPACEXTBL where Landing_Outcome = "Success (ground pad)";

min(DATE)
2015-12-22

Using the function MIN works out the minimum date in the column Date

The WHERE clause filters the dataset to only perform calculations on Landing_Outcome Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL query : select Customer, BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME ='Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 and 6000;

Customer	Booster_Version	PAYLOAD_MASS__KG_
SKY Perfect JSAT Group	F9 FT B1022	4696
SKY Perfect JSAT Group	F9 FT B1026	4600
SES	F9 FT B1021.2	5300
SES EchoStar	F9 FT B1031.2	5200

The WHERE clause filters the dataset to Landing_Outcome = Success (drone ship)

The AND clause specifies additional filter conditions

Payload_MASS_KG_>4000ANDPayload_MASS_KG_<6000

Total Number of Successful and Failure Mission Outcomes

SQL query : select MISSION_OUTCOME, count(*) from SPACEXTBL GROUP BY MISSION_OUTCOME;

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

- There are 1 failure in flight, 99 successes and 1 success with unclear payload status.

Boosters Carried Maximum Payload

SQL query : select BOOSTER_VERSION, Max_Payload from (Select BOOSTER_VERSION, MAX(PAYLOAD_MASS__KG_) as Max_Payload from SPACEXTBL Group BY BOOSTER_VERSION);

BOOSTER_VERSION	Max_Payload
F9 B4 B1039.2	2647
F9 B4 B1040.2	5384
F9 B4 B1041.2	9600
F9 B4 B1043.2	6460
F9 B4 B1039.1	3310
F9 B4 B1040.1	4990
F9 B4 B1041.1	9600
F9 B4 B1042.1	3500
F9 B4 B1043.1	5000
F9 B4 B1044	6092

Different booster version has different max payload mass.

2015 Launch Records

SQL query : SELECT DATE,Mission_Outcome,Booster_Version,Launch_Site from SPACEXTBL where Date LIKE '%2015%';

Date	Mission_Outcome	Booster_Version	Launch_Site
2015-01-10	Success	F9 v1.1 B1012	CCAFS LC-40
2015-02-11	Success	F9 v1.1 B1013	CCAFS LC-40
2015-03-02	Success	F9 v1.1 B1014	CCAFS LC-40
2015-04-14	Success	F9 v1.1 B1015	CCAFS LC-40
2015-04-27	Success	F9 v1.1 B1016	CCAFS LC-40
2015-06-28	Failure (in flight)	F9 v1.1 B1018	CCAFS LC-40
2015-12-22	Success	F9 FT B1019	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

SQL query : SELECT LANDING_OUTCOME, count(*) FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' Group by LANDING_OUTCOME ORDER BY DATE DESC;

Function COUNT counts records in column WHERE filtres data LIKE (wildcard) AND (conditions)AND (conditions)

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

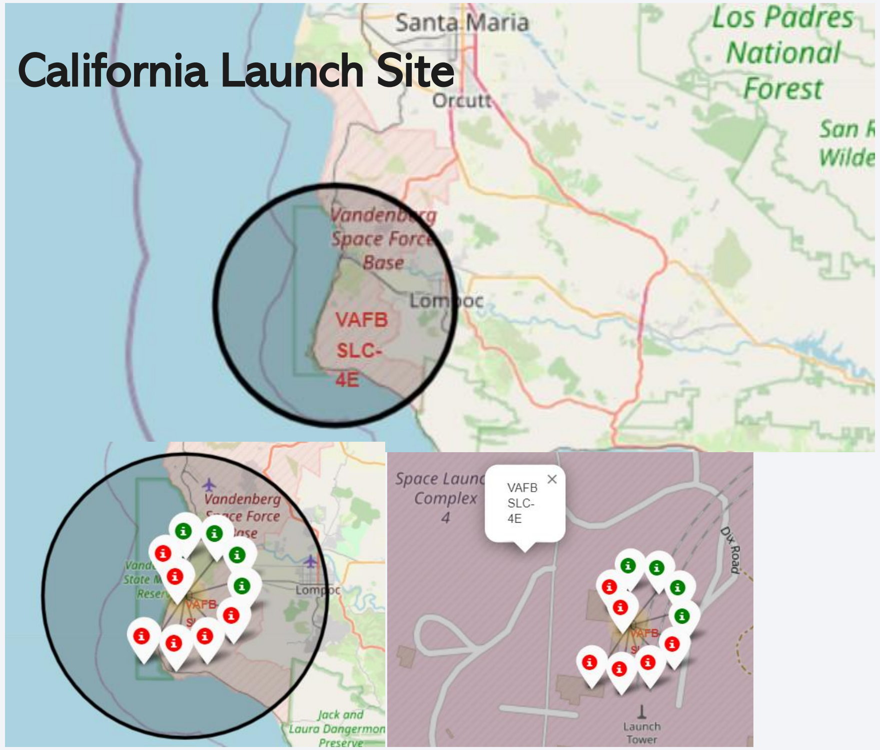
Section 3

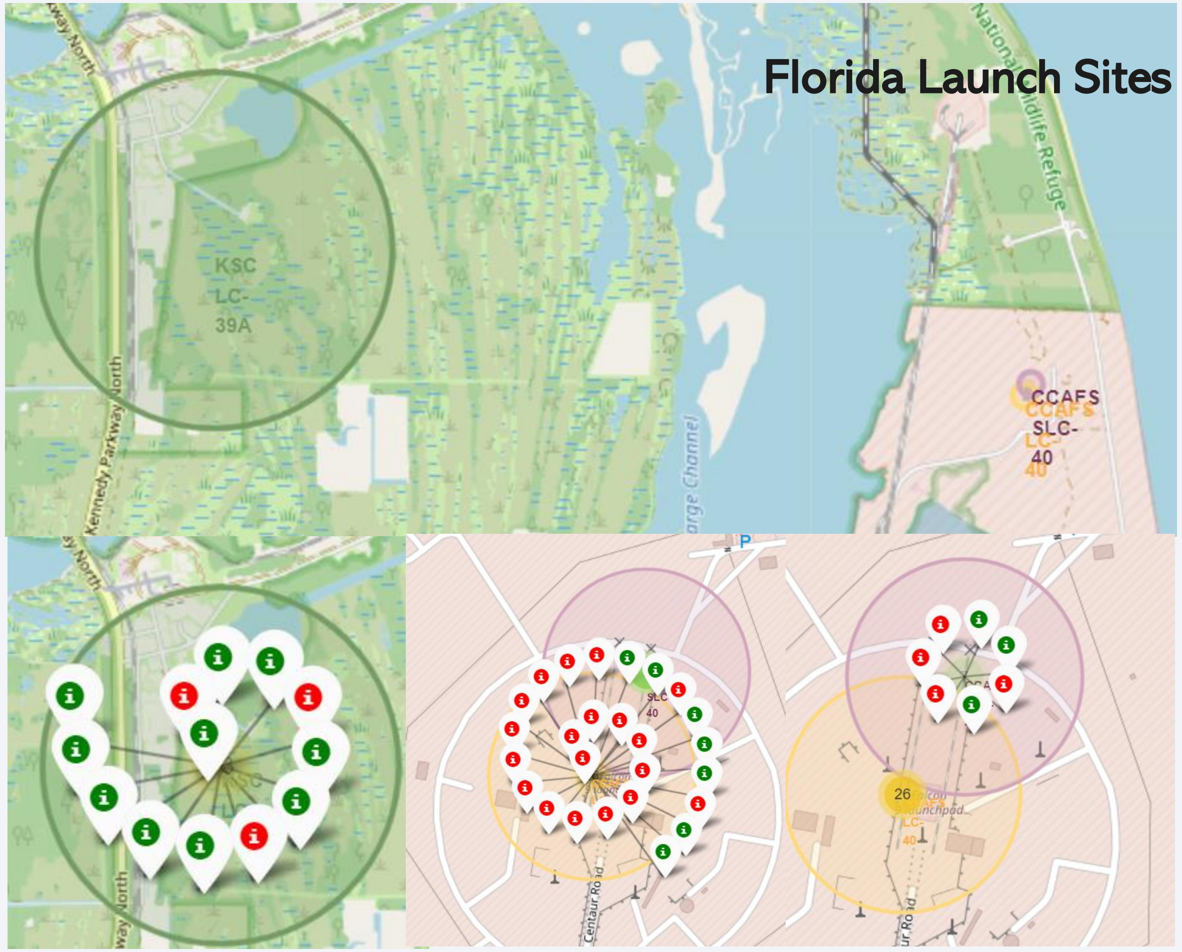
Launch Sites Proximities Analysis

All launch sites global map markers

We can see that the SpaceX launch sites are in the United States of America coasts, Florida and California

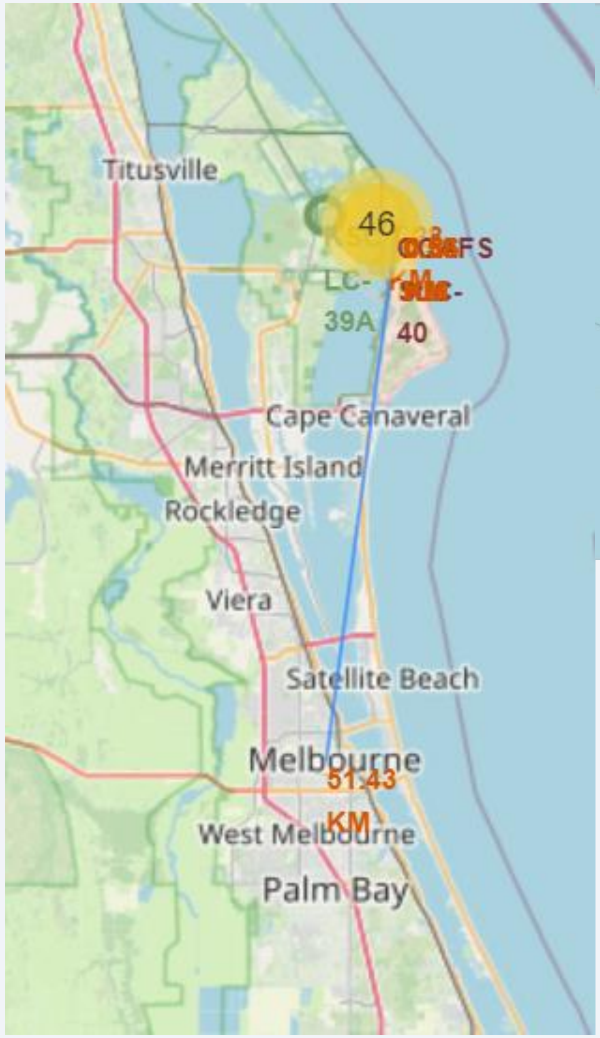
Colour Labelled Markers

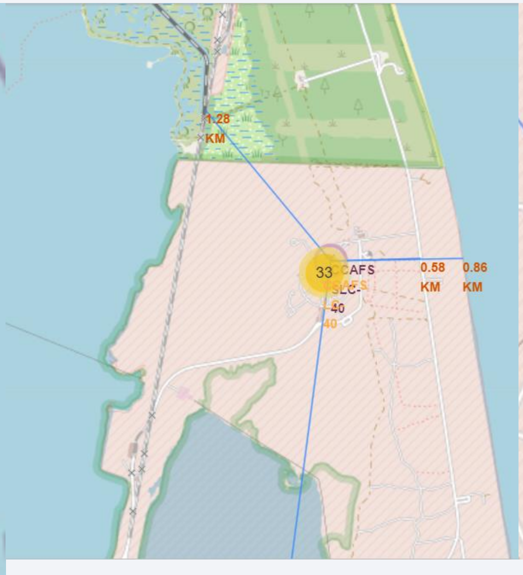




Green Marker shows successful Launches andRed Markershows Failures

The proximity of the launch sites





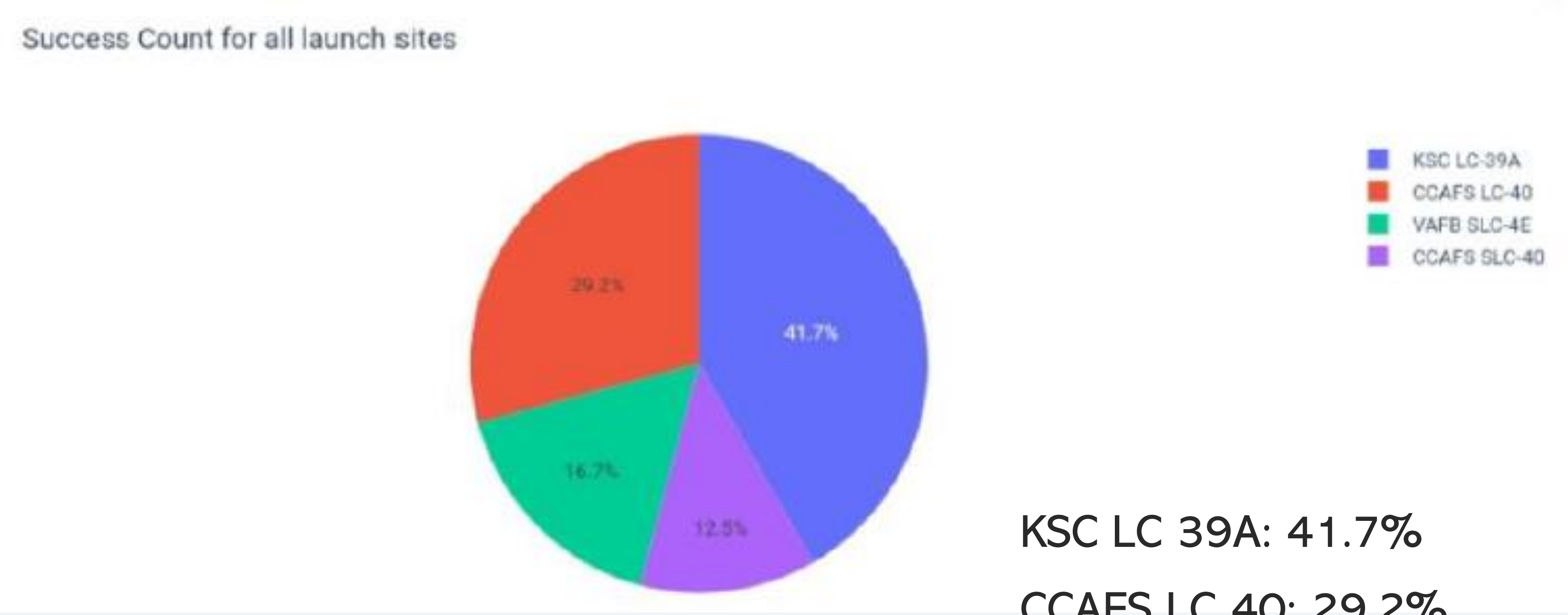
CCAFS LC-40 is :

- 0.58 km from the Street “Samuel C Phillips”,
- 0.86 km from the coast
- 1.28 km from the railway
- 51.43 km from Melbourne

Section 4

Build a Dashboard with Plotly Dash

The Success Percentage Achieved by Each Launch Site



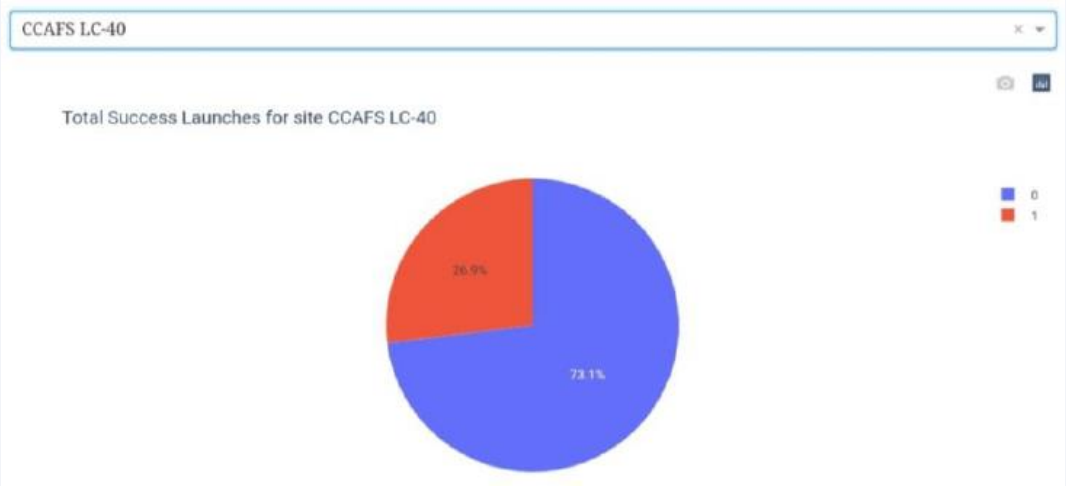
KSC LC 39A: 41.7%

CCAFS LC 40: 29.2%

VAFB SLC 4E: 16.7%

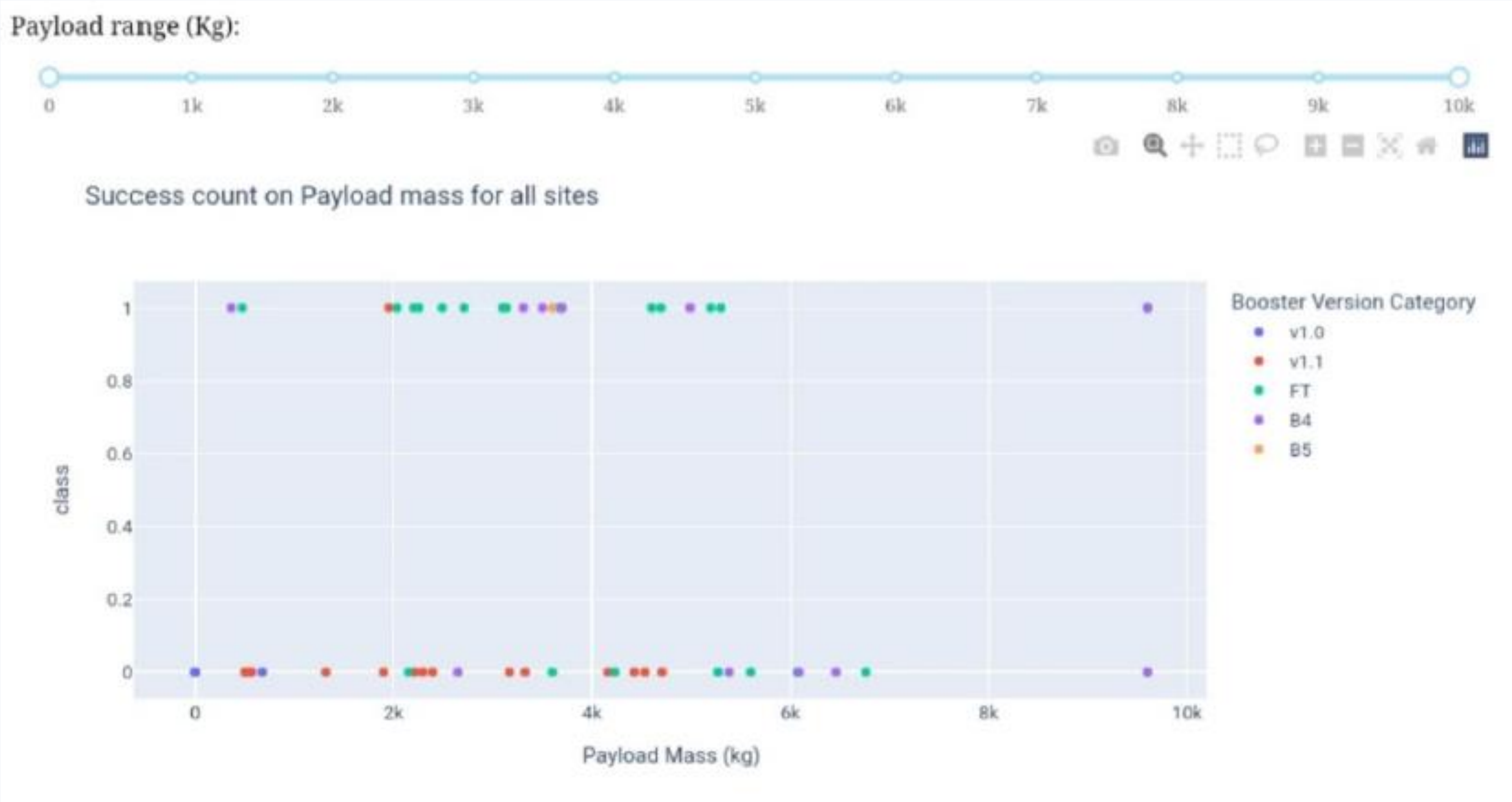
CCAFS SLC 40: 12.5%

Launch site with highest launch success ratio



KSC LC 39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs Launch Outcome

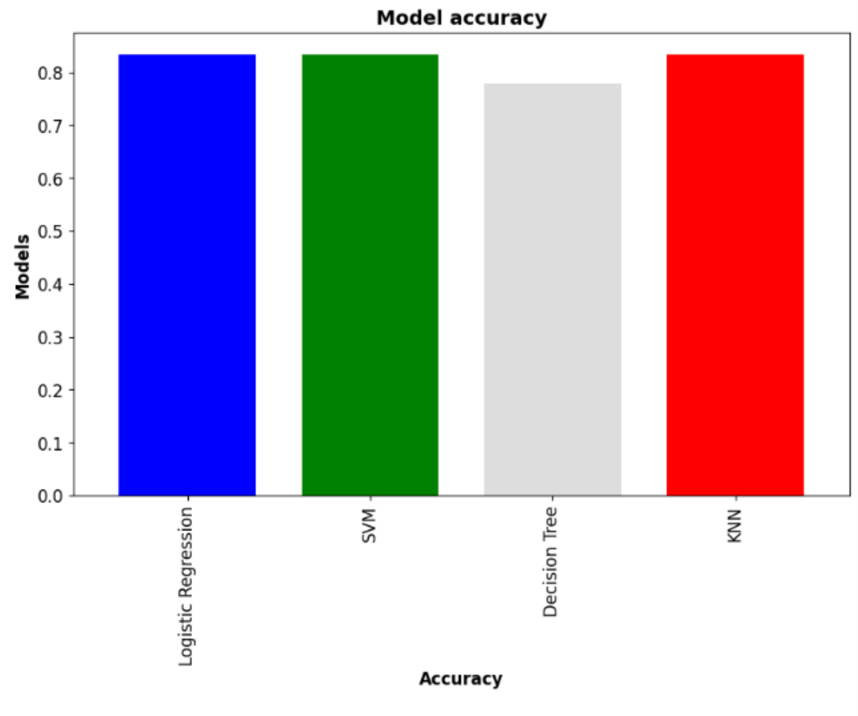


Section 5

Predictive Analysis (Classification)

Classification Accuracy

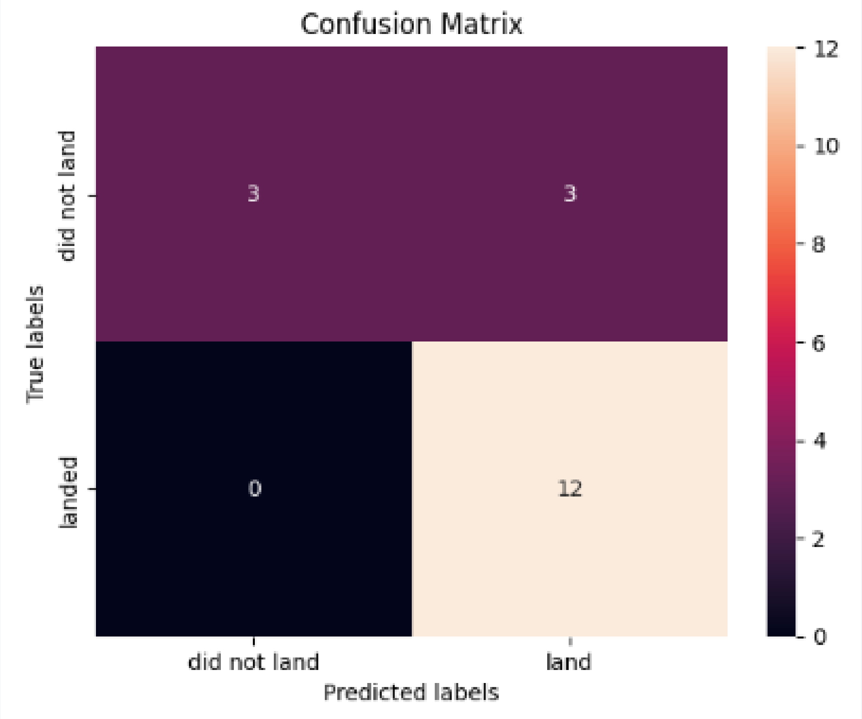
Model	Accuracy
Logistic Regression	0.833333
SVM	0.833333
Decision Tree	0.777778
KNN	0.833333



After obtaining the accuracy of our various models, for the decision tree classifier using the validation data, we achieved 78% accuracy on the test data.

Confusion Matrix for Decision Tree

- Examining the confusion matrix, we see that Tree can distinguish between the different classes.
- The predictive model tells us that there will be 3 true positive, 12 true negative, 0 false positive and 3 false negative .



Conclusions

1. Launch site choice significantly influences success rates, and there is a correlation between payload mass and success, with heavier payloads decreasing the likelihood of the first stage's return.

2. Specific orbit types, such as ES-L1, GEO, HEO, and SSO, exhibit the highest success rates, contrasting with SO, which has the least success.

3. The overall success rate has witnessed a consistent increase since 2013, reaching its peak in 2020, indicating a positive trend over the years.

4. Utilizing the decision tree classifier with optimal parameters yielded the highest prediction accuracy at 78%, establishing it as the most effective algorithm for this dataset.

5. The performance of low-weighted payloads surpasses that of heavier payloads, suggesting a noteworthy impact on success rates based on payload mass.

Appendix

- https://github.com/CesarCeballos0126/Space-X-Falcon-9-First-Stage-Landing-Prediction

Thank you!