IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Paola ALLEGRINI
23/01/2025

Outline

Executive Summary

- Data collection and Wrangling of past Falcon 9 launches
- Exploratory Data Analysis
- Model selection and training to predict First Stage success to land
- Results:
  - We were able to predict with a 83% accuracy the outcome of the first stage landing of a rocket

Introduction

- *This project is conducted in the context of the Data Science IBM Coursera Course*

- SpaceX advertises Falcon 9 rockets launches with a cost of 62 million dollars. Much of the savings is because SpaceX can reuse the first stage (one of the most expensive parts).
- Our objected is, analyzing data from previous launches, predict if the first stage will land and therefore be reused.
- Predicting the landing outcome we allows us to determine the cost of a launch.
- Insights: Useful information if alternate company wants to bid against SpaceX for a rocket launch

Section 1

Methodology

Methodology

Executive Summary

1. Data collection methodology
2. Perform data wrangling
3. Perform exploratory data analysis (EDA) using visualization and SQL
4. Perform interactive visual analytics using Folium and Plotly Dash
5. Perform predictive analysis using classification models

Data Collection

The data was collected from two sources:

1. From the SpaceX API:

   historical data for launches between 2006 and 2020 (API)

2. From Webscraping:

   historical launch records (Wikipedia page)

From each source we retrieved valuable information on the launches including:

- BoosterVersion
- Launchsite
- Orbit
- PayloadMass
- Launch Outcome …

Data Collection – SpaceX API

- Requested and parsed the SpaceX launch data using the GET request
- Defined specific get functions to collect each feature using their identification number
- Filtered data frame to only include Falcon 9 launches
- Replaced missing values in PayloadMass by the average

- notebook

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

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

```
# Takes the dataset and uses the rocket column to call the API and append the data to the li
def getBoosterVersion(data):
    for x in data['rocket']:
       if x:
        response = requests.get("https://api.spacexdata.com/v4/rockets/"+str(x)).json()
        BoosterVersion.append(response['name'])
```

	rocket	payloads	launchpad	cores	flight_num
0	5e9d0d95eda69955f709d1eb	5eb0e4b5b6c3bb0006eeb1e1	5e9e4502f5090995de566f86	{'core': '5e9e289df35918033d3b2623', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	
1	5e9d0d95eda69955f709d1eb	5eb0e4b6b6c3bb0006eeb1e2	5e9e4502f5090995de566f86	{'core': '5e9e289ef35918416a3b2624', 'flight': 1, 'gridfins': False, 'legs': False, 'reused': False, 'landing_attempt': False, 'landing_success': None, 'landing_type': None, 'landpad': None}	

Data Collection - Scraping

- Used BeautifulSoup() to parse the html table containing launch records
- Extracted column names and content
- notebook

```python
# Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(response.content, "html.parser")
#soup = BeautifulSoup(response.text, "html.parser")
```

```python
column_names=[]
for col in first_launch_table.find_all('th'):
    col_name=extract_column_from_header(col)
    if col_name and len(col_name)>0:
        column_names.append(col_name)

column_names
```

	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 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44

Data Wrangling

- From the API data
- Explored Launch Sites and Orbits
- Explored Landing Outcomes:
  - 8 Types of Outcome
  - 5 bad outcomes
  - 3 good outcomes
- Added a label ‘Class’
  - 1: Successful landing (good outcome)
  - 0: Failed landing (bad outcome)

```
for i,outcome in enumerate(landing_outcomes.keys()):
    print(i,outcome)
```

```
0 True ASDS
1 None None
2 True RTLS
3 False ASDS
4 True Ocean
5 False Ocean
6 None ASDS
7 False RTLS
```

	Outcome	Class
0	None None	0
1	None None	0
2	None None	0
3	False Ocean	0
4	None None	0
5	None None	0
6	True Ocean	1
7	True Ocean	1

- [notebook]

EDA with Data Visualization & SQL

With Data Visualisation

- Compared various variables with each other to observe if there was a relationship:
  - Flight number vs Payload
  - Payload vs Orbit Type …
- Aimed to find what features influence the Landing outcome
  - Success Rate vs Orbit Type …
- notebook

With SQL

- Performed SQL queries to better understand the SpaceX dataset:
  - Queries on Launch sites
  - Average Payload Mass
  - First Launches …
- notebook

Build an Interactive Map with Folium

- Map to gain insights on how the position of a launch site could influence the outcome of a Landing and how it is chosen
- Added a marker for each launch site
- For each launch site:
  - red marker: Failure to land
  - green marker: Success to land
- Calculated the distance of the closest highway, railway and city to a launch site
- notebook

Build a Dashboard with Plotly Dash

- Built a dashboard displaying:
  - A pie chart with the ratio of successful launches by site
  - A pie chart proportion of Success/Failure to land for each site
  - Scatter plot of Success Launch vs Payload Mass with booster version for each launch
- The goal was to gain insights on the influence of the Launch site, Payload Mass and Booster Version in the Success of the first stage to land
- notebook

Predictive Analysis (Classification)

- Trained 4 classification models to predict the Success/Failure to land of a launch.
- The 4 classification models where:
    - Logistic Regression
    - SVM
    - Decision tree classifier
    - K Nearest Neighbors

- For each of them:
    - I split the dataset into a train(80%) and a test set (20%)
    - Trained the models and selected the best parameters using GridSearchCV
    - Tested and validated the models :
        - Displayed their confusion matrix
        - Calculated their accuracy score
- notebook

Results

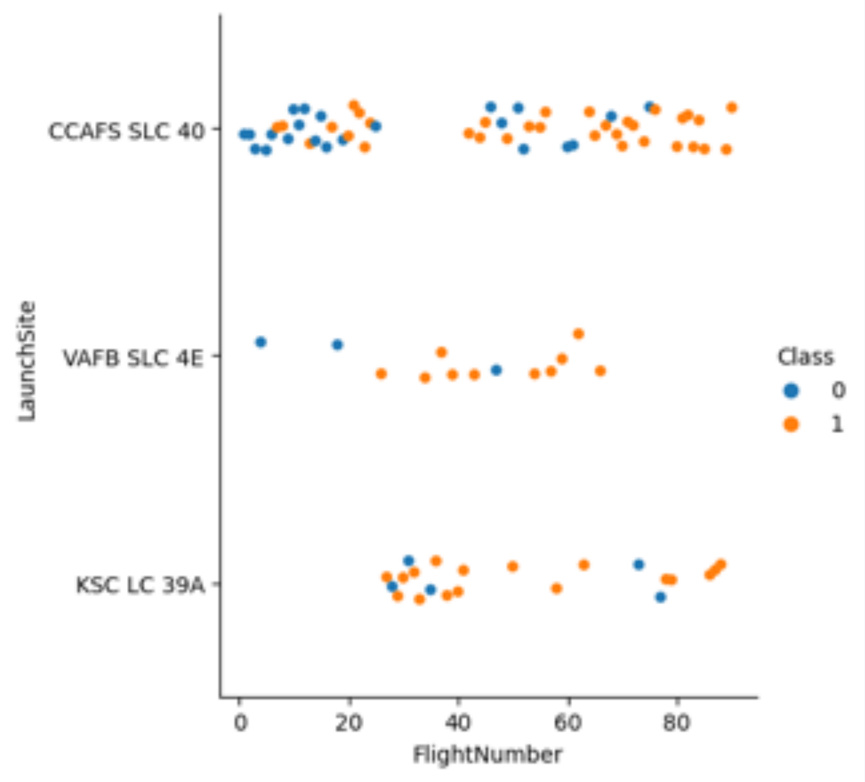
- Exploratory data analysis results:
  - Successful landing increases with number of flights
  - Payload Mass on a launch depends on orbit type and Launch site

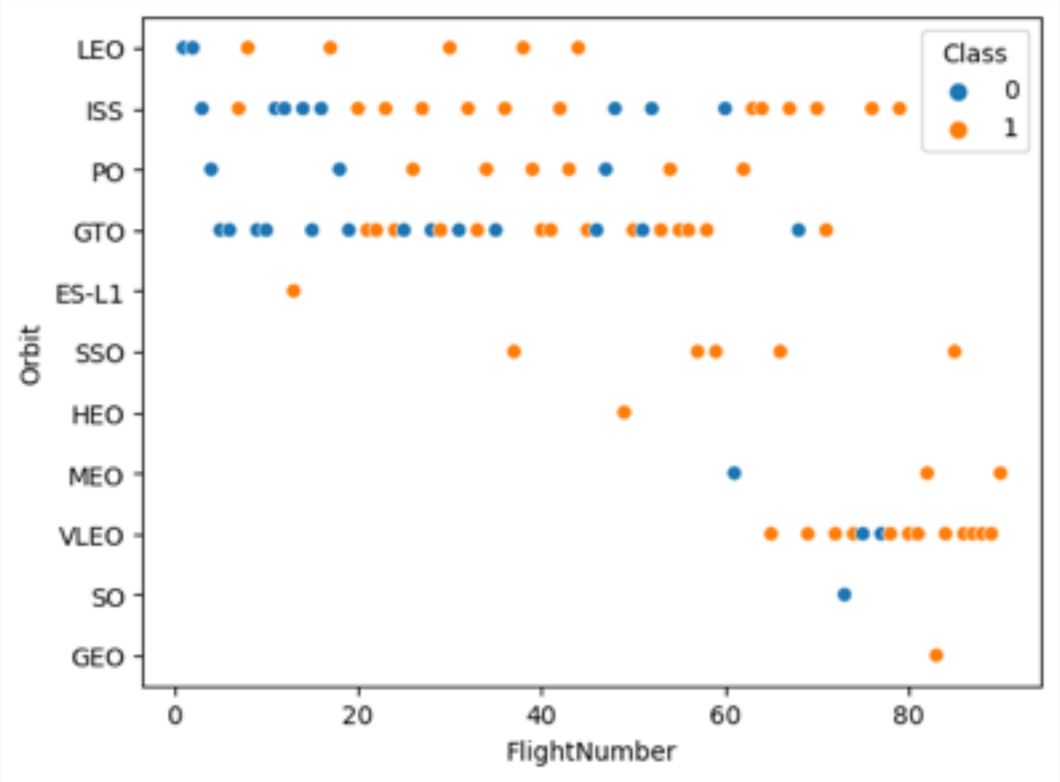
- Predictive analysis results
  - We were able to predict the success landing of a launch with 83.33% accuracy

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site & Orbit Type

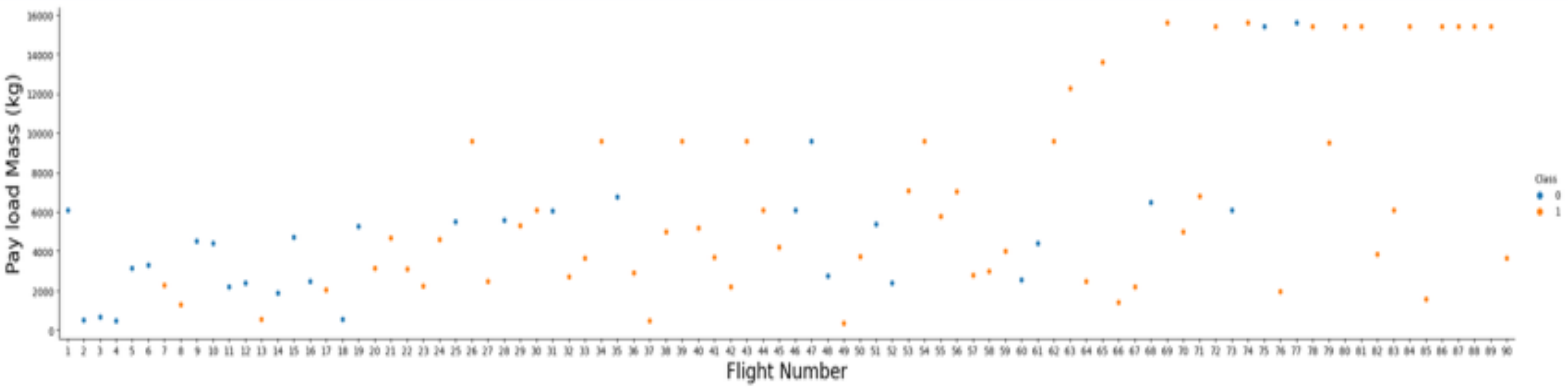




- It seems LaunchSite VAFB SLC got discontinued
- The most used Launch Site is CCAFS SLC 40
- It is unclear which Launch Site has a highest success rate

- In the LEO orbit, success seems to be related to the number of flights.
- he GTO orbit, there appears to be no relationship between flight number and success.

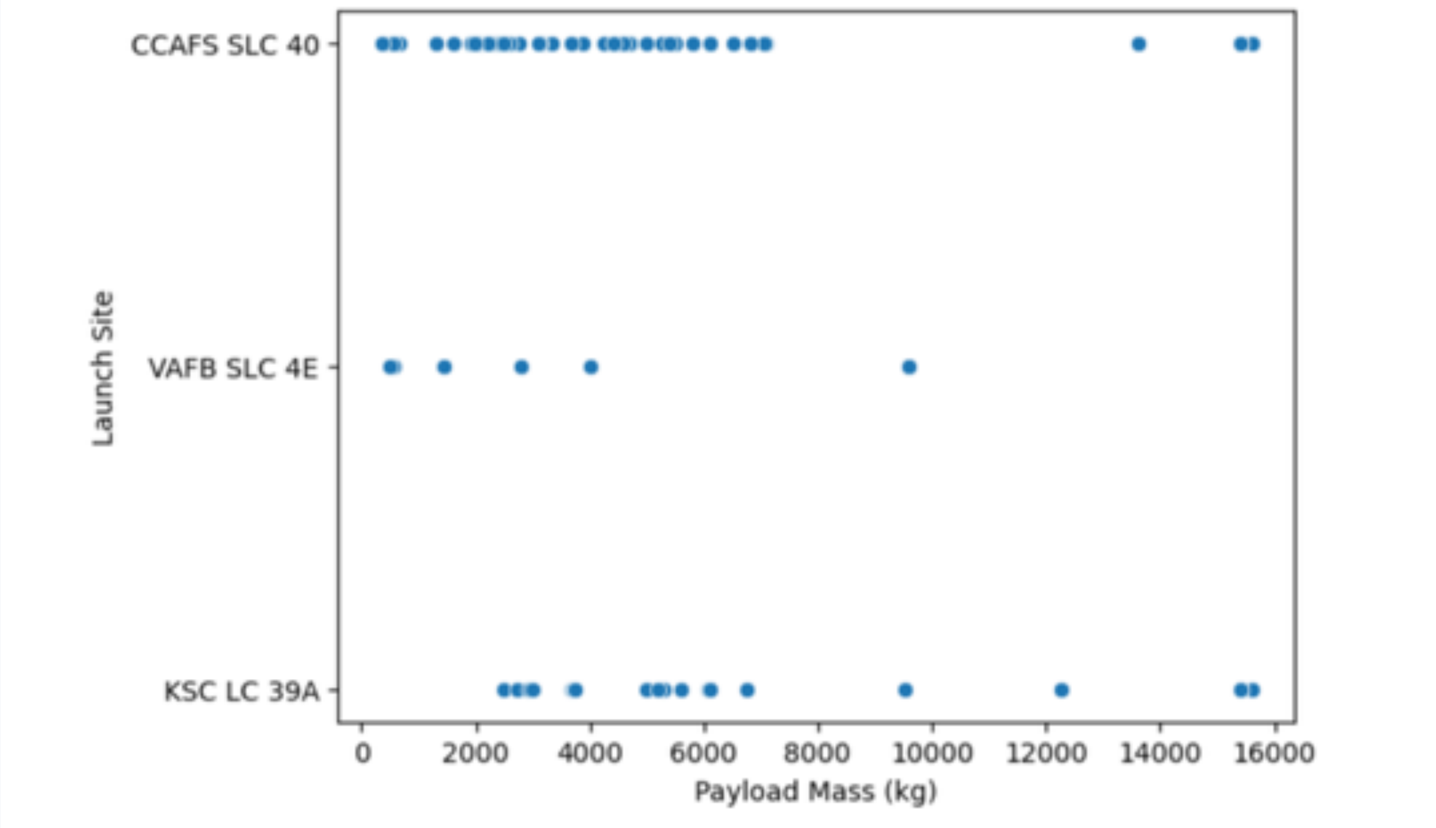
Payload vs. Flight Number



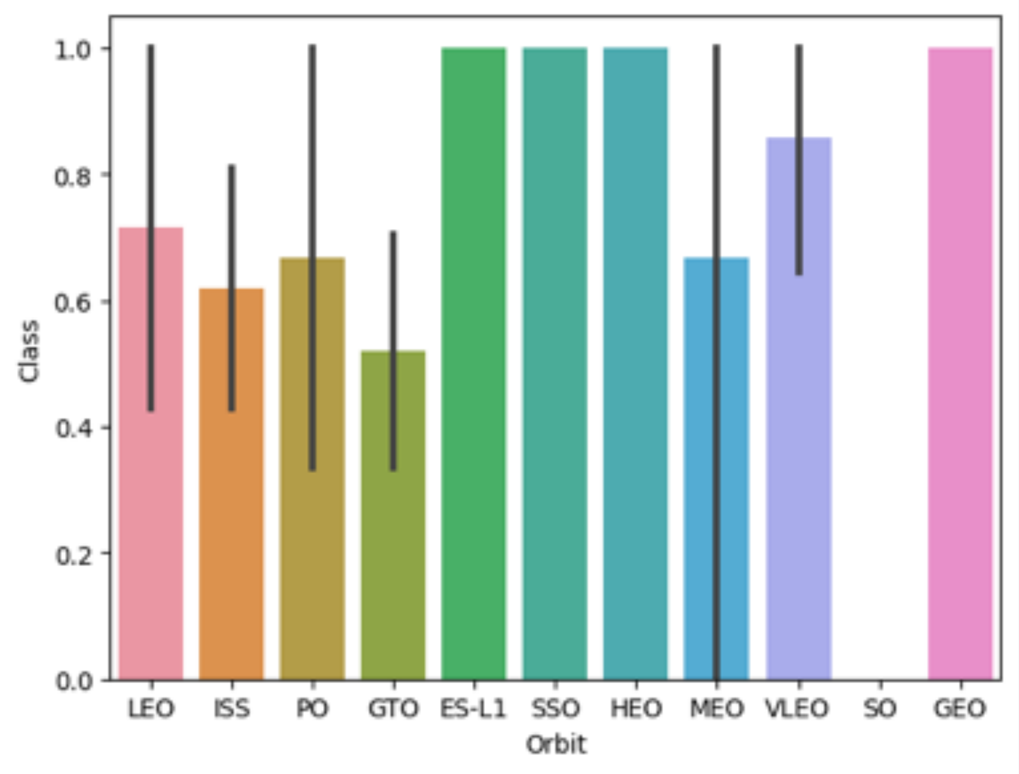
- With time, launches used heavier Payloads
- The success rate has grown with the number of flights

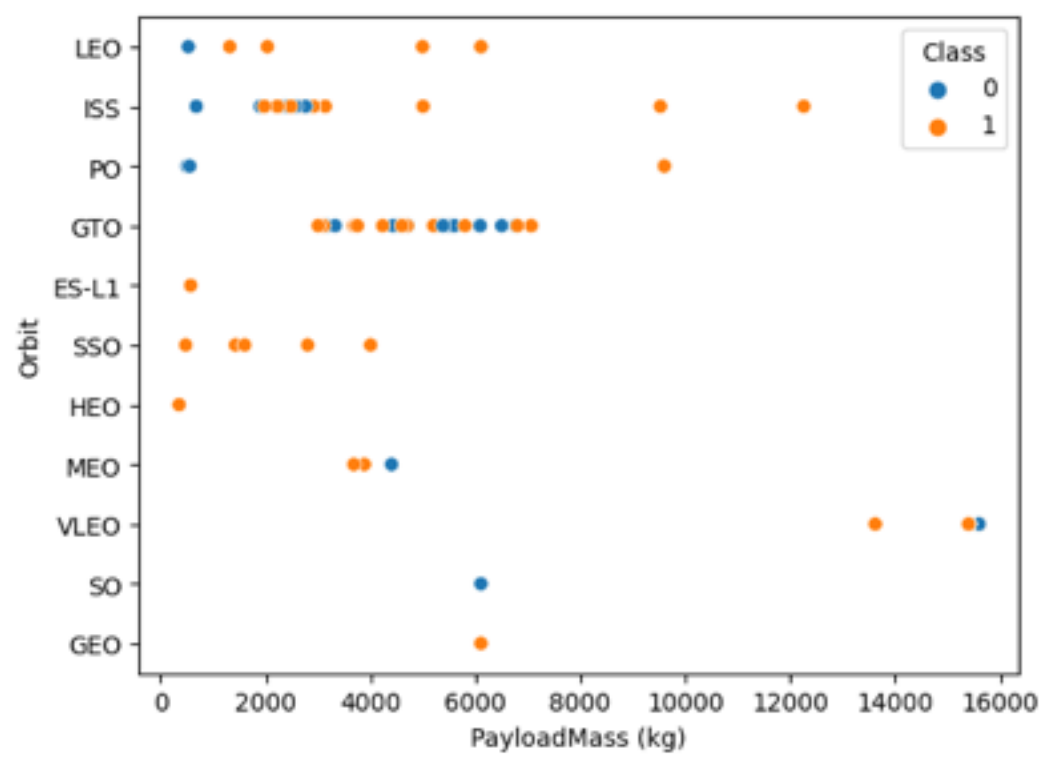
Launch Site vs.Payload Mass

- Launch site VAFB SLC 4E does not launch rockets heavier that 10’000 kg



Success Rate & Payload vs. Orbit Type



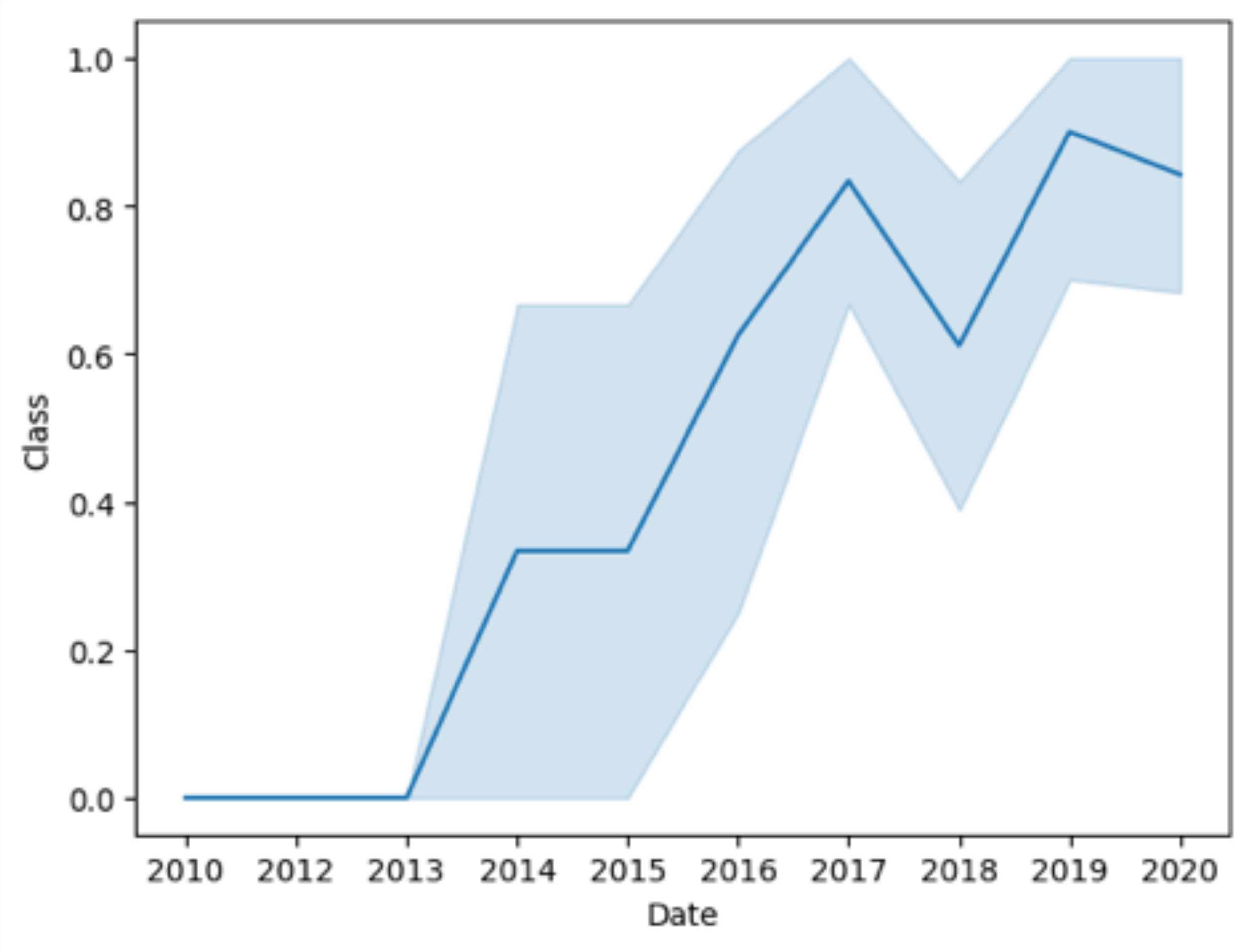


- ES-L1, SSO, HEO and GEO have the highest success rate

- The only Orbit types dealing with heavy rockets (>10,000kg)are ISS and VLEO
- For GTO the Payload does not seem to be related to the success rate

Launch Success Yearly Trend

- Since 2013, the launch success rate has kept growing



Insights on the SpaceX dataset

- They are 4 different launch sites

```
%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

Launch Site Names Begin with 'CCA'

- The first 5 records where launch sites begin with `CCA` are between 2010 and 2013

```
%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
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO
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)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)

Payload Mass

- Total payload carried by boosters from NASA : 4,5596 kg

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_)
45596

- The average payload mass by F9 v1.1 : 2,337.8 kg

```
%sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1 %';
```

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

avg(PAYLOAD_MASS__KG_)
2337.8

First Successful Ground Landing Date

- The first successful landing outcome on ground pad dates from 2015-12-22

```
%sql select MIN(DATE) from SPACEXTABLE where Landing_Outcome like 'Success (ground pad)';
```

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

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- 4 different booster versions have successfully landed on drone ship and had 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

Total Number of Successful and Failure Mission Outcomes

- On 101 missions, only 1 failed

success_count	failure_count
100	1

Boosters Carried Maximum Payload

- List 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 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

2015 Launch Records

- In 2015, two launches have failed to land in drone ship

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

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

- Landing outcomes between the date 2010-06-04 and 2017-03-20

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

Section 3

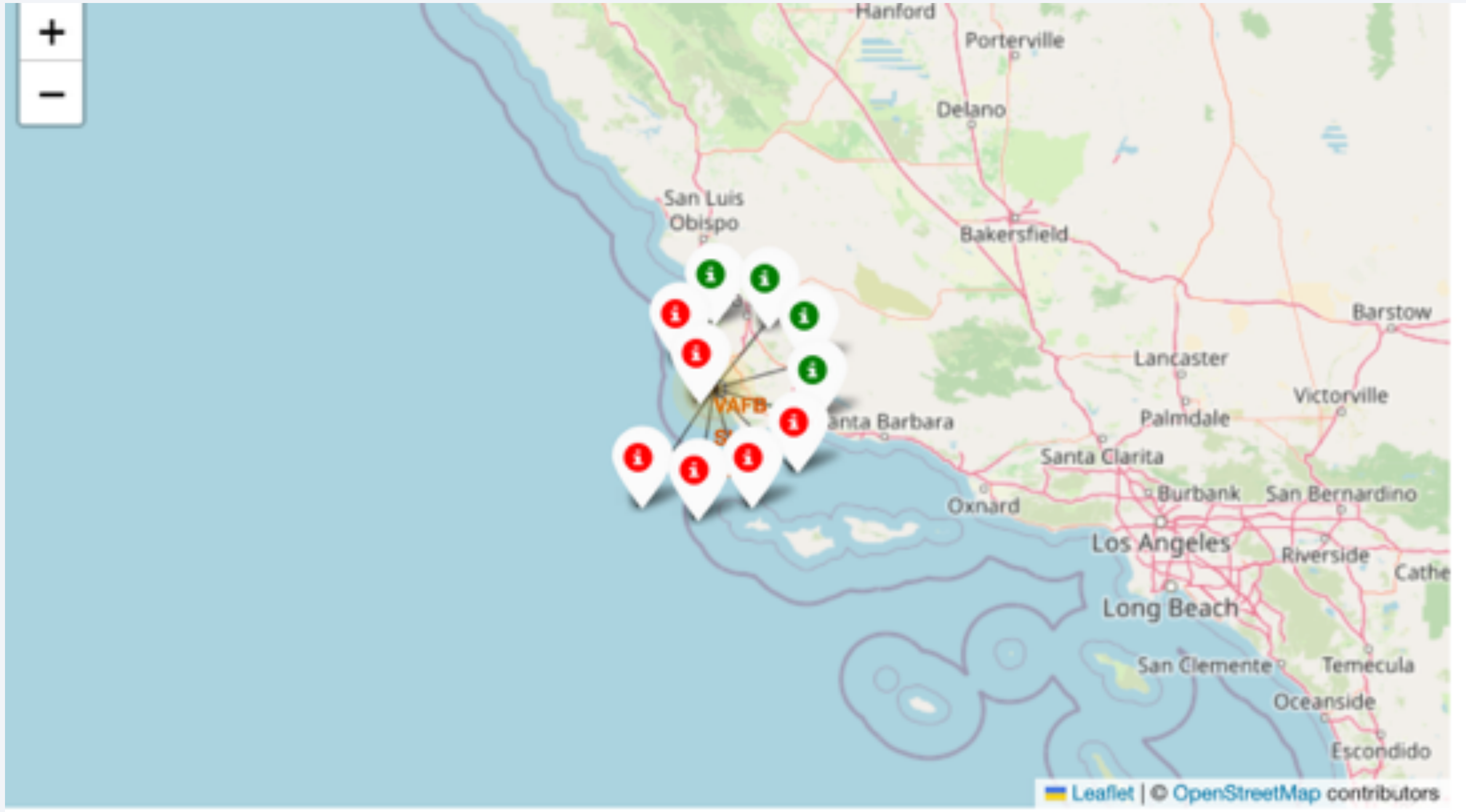
Launch Sites Proximities Analysis

Launch Sites positions

- All launch sites are situated in the South of the US, close to the Equator line
- VAFB SLC-4E is situated on the east coastline and the others on the east coastline.

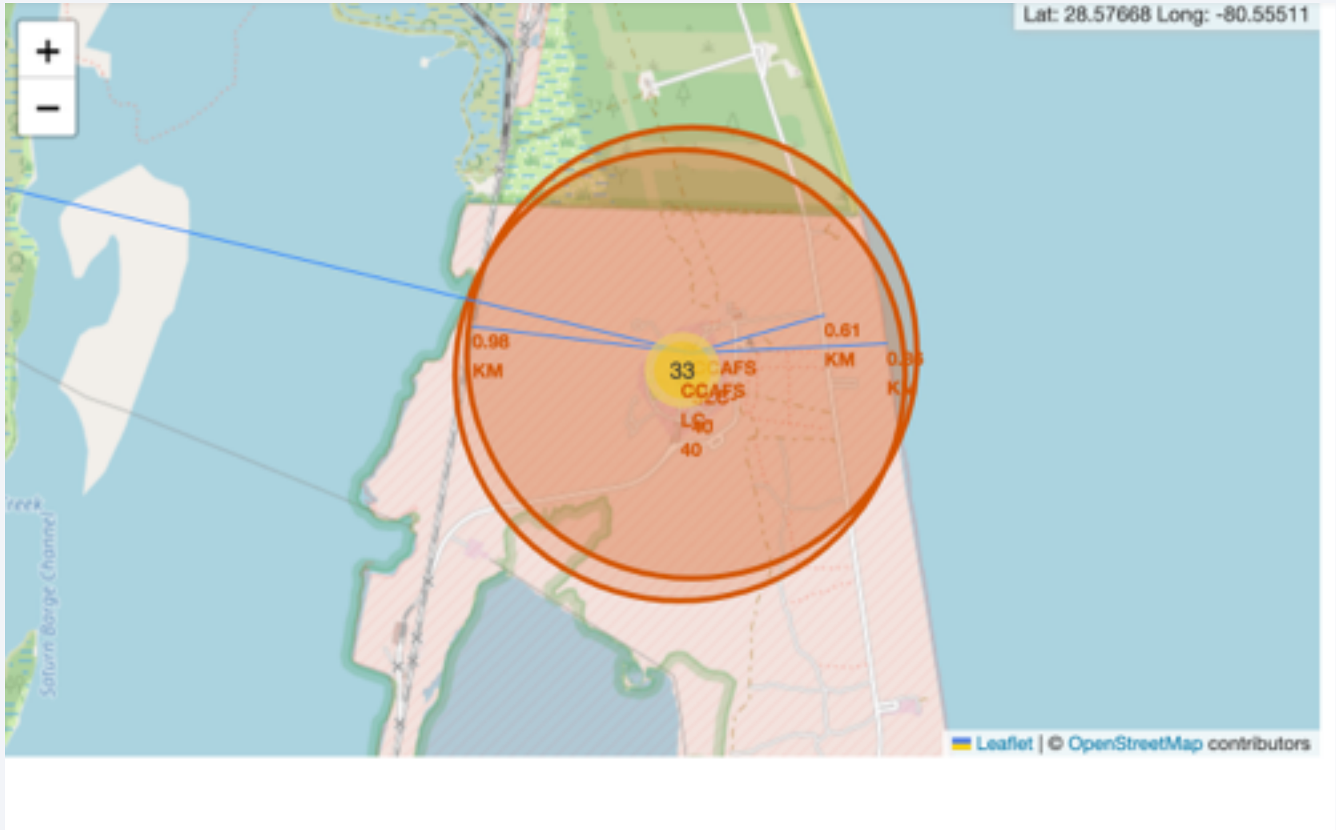
Landing outcomes for Launch Site VAFB SLC 4E

- 10 launches
- More failures to land than successes.



Proximities of a Launch Site

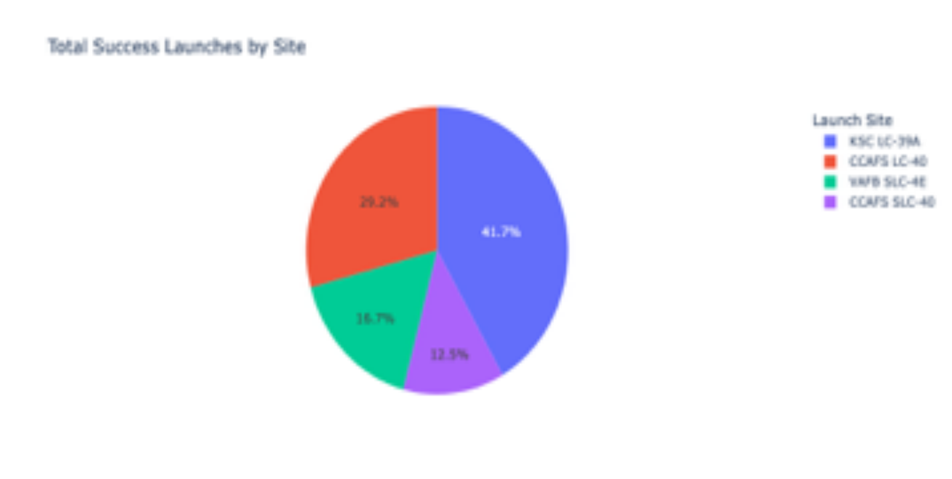
- The launch sites are situated :
  - close to the coastline
  - not far from railways and highways
- They keep a distance with cities

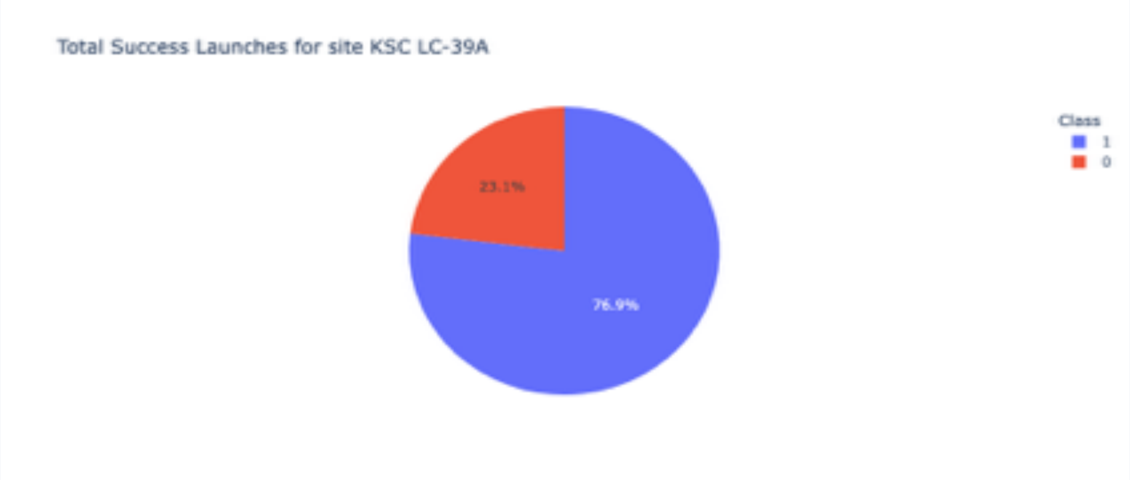


Section 4

Build a Dashboard with Plotly Dash

Total Success Launches

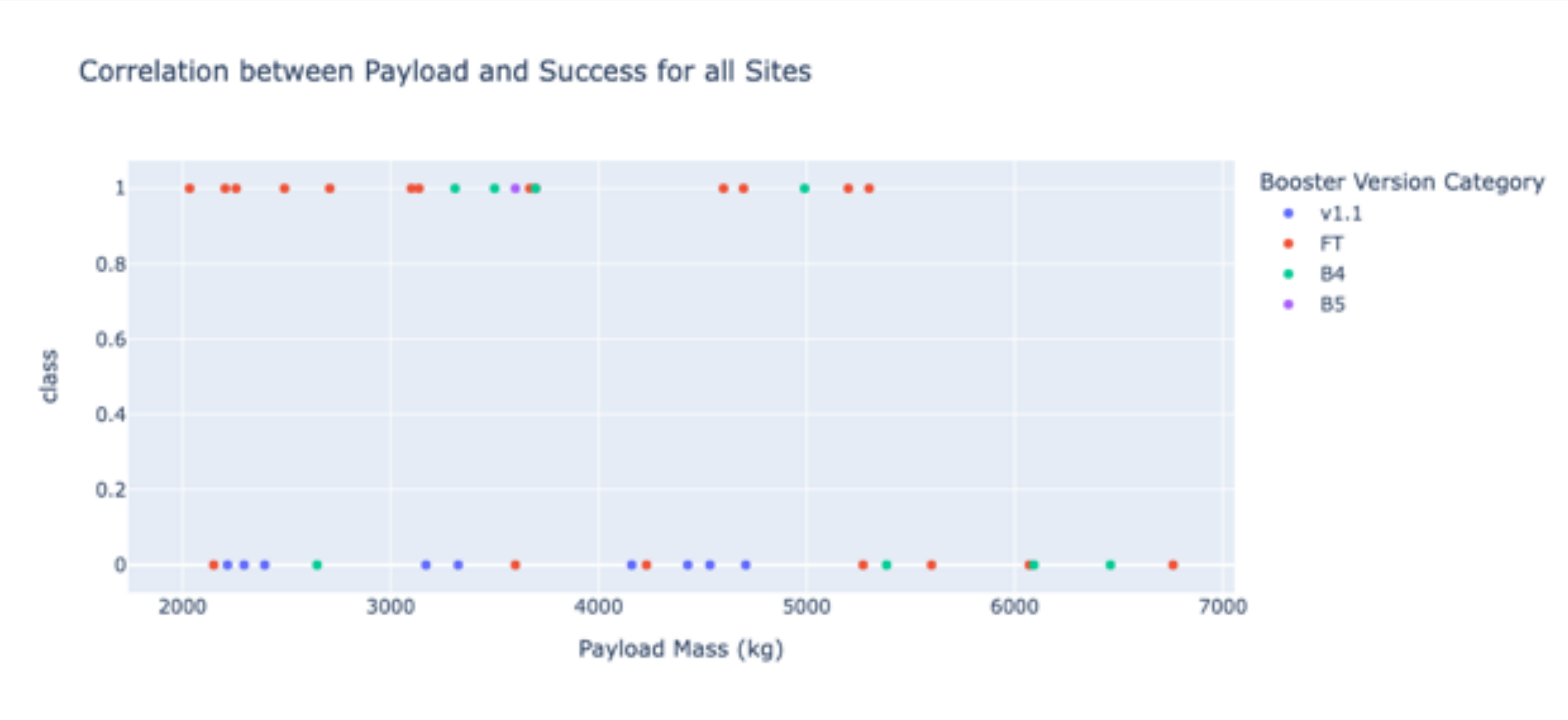




- Among the 4 sites:
  - KSL LC-39 A has the highest number of successful launches
  - CCAFS SLC-40 has the least

- For KSC LC -39A:
  - More than 75% of the launches have been successful.

Payload vs Launch Outcome



- All the launches with payload mass between 6,000 and 7,000 kg have failed to land
- Launches for Booster version v1.1 and Payload mass between 2,000 and 5,000 kg have failed to land
- Most of the successes in this payload mass range are with Booster Version FT

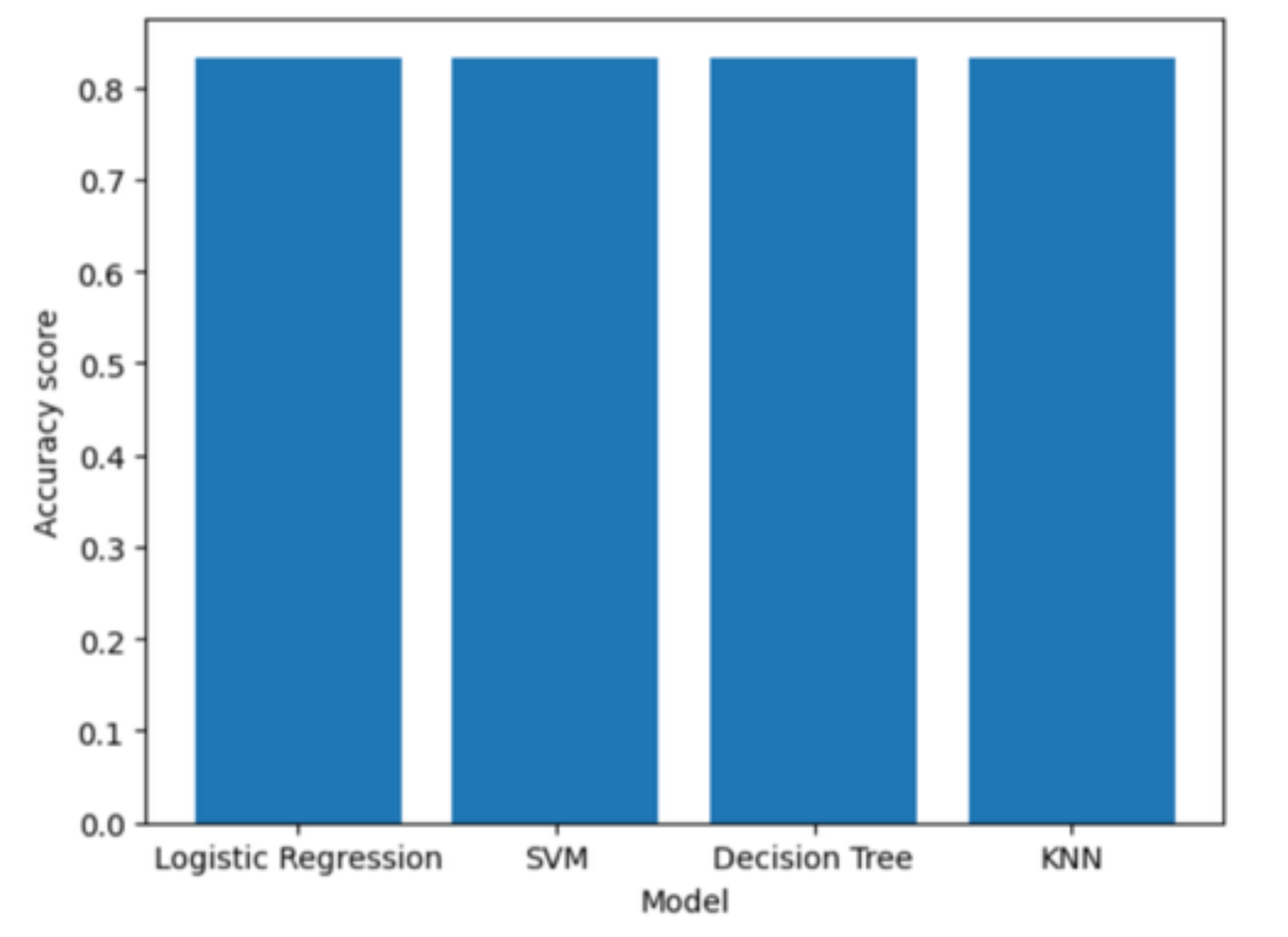
Section 5

Predictive Analysis (Classification)

Classification Accuracy

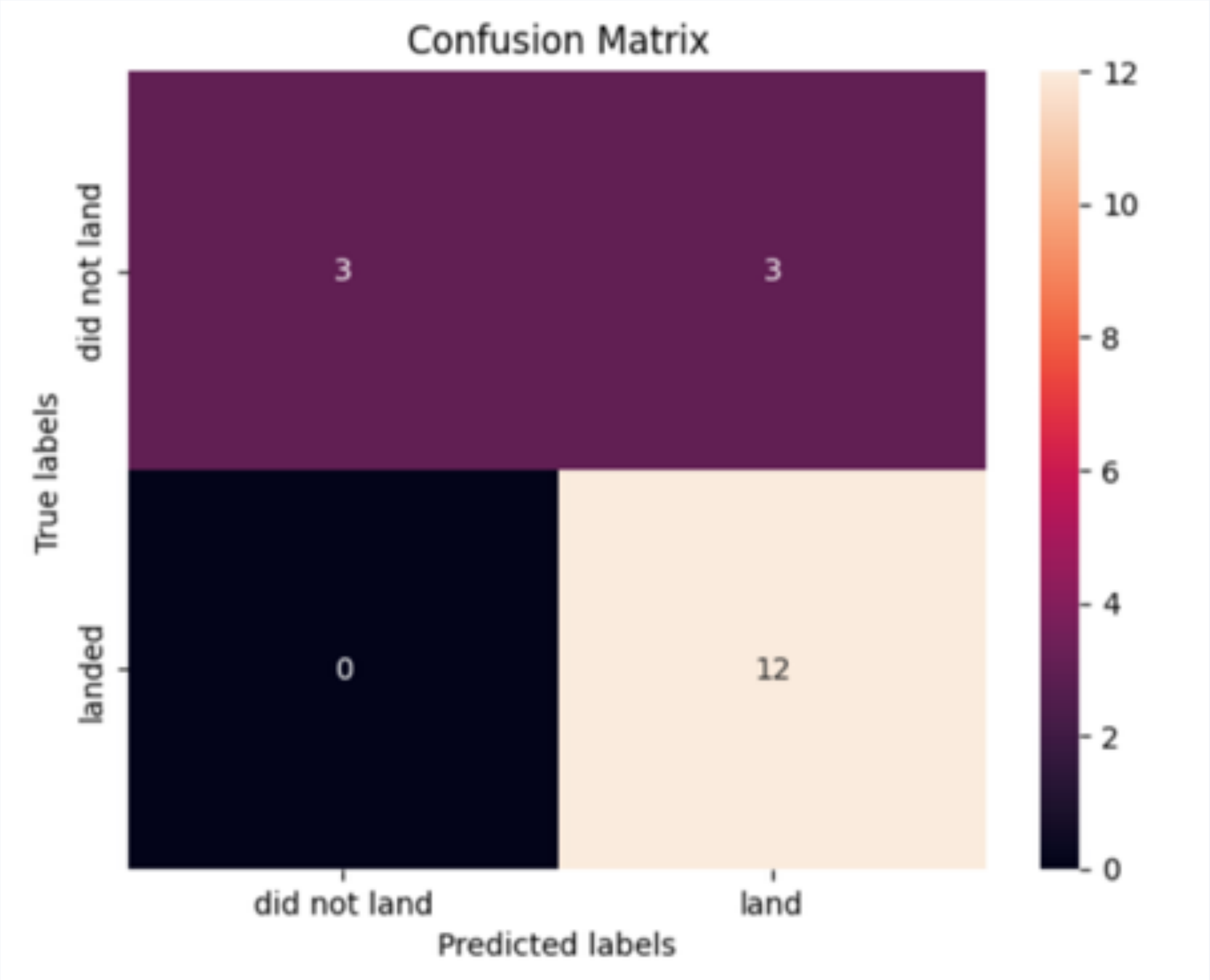
- We trained and tuned the parameters of classification model:
    - Logistic Regression
    - SVM
    - Decision Tree Classifier
    - KNN

- All performed equivalently with a 83.33% accuracy on the testing dataset



Confusion Matrix

- For the four models, the confusion matrix is the following
- They correctly label the successful landed
- There are 3 False Positive, meaning they do not always correctly label the failures to land.



Conclusions

- We were able to predict with 83.33% accuracy if the first stage of a launch will land.
- This information will be useful to estimate the cost of the SpaceX launch bid against it.

Thank you!