IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alvaro Ortiz de Pazos
18 of October 2021

Outline

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

Executive Summary

Due the development in technology in the last 20 years, new companies have appear in the race to travel to the space.

One of this companies is SpaceX by Elon Musk, and they are trying to do it by minimizing the cost as much as possible, due the main issue of travelling to the space is the cost of it. The launch of a rocket can cost more than 165 millions dollars.

In this projects I have use data science and machine learning in order to try to predict if a rocket will be able to land in the earth after it launch within a cost reduction of the rocket.

Introduction

- Elon Musk company, SpaceX is trying to reuse rockets by trying to land them after they launch. Therefore, I have use historial data of Space X in order to create a model with machine learning that will help me to predict if the landing will be successful or not.

- In this project we will try to answer some questions, some of them are:
  - What key factor impact the outcome of a landing?
  - How can we plan more successful experiments?
  - Should an experiment be attempted under set conditions?

Cost of 1 rocket launch +165 million dollars

Tourism market estimated to be worth 3$ billion by 2030

SpaceX estimates to save 62% with rocket use

Section 1

Methodology

Methodology

Executive Summary

Data collection methodology:

- Different methods were used to collect the data like, SpaceX API calls or Web Scraping

Perform data wrangling

- Calculation of significant metrics such as averages and success rates, and cleansing of messy data

Perform exploratory data analysis (EDA) using visualization and SQL

Perform interactive visual analytics using Folium and Plotly Dash

Perform predictive analysis using classification models

- Use of splitting data and test sets, multiple classification models, optimized hyperparameters, and validated the accuracy of the models
  - SVM, KNN, Decision Tree and Logistic Regresion

Data Collection – SpaceX API

API GET Request

JSON Object

Pandas DataFrame

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

```
response_static = requests.get(static_json_url)
```

We should see that the request was successfull with the 200 status response code

```
response.status_code
```

```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
data = pd.json_normalize(response_static.json())
```

Using the dataframe `data` print the first 5 rows

```
# Get the head of the dataframe
data.head()
```

https://github.com/AlvaroOrtiz2001/spacex-project/blob/main/jupyter-labs-spacex-data-collection-api%20(FINAL)%20(1).ipynb

Data Collection - Scraping

- URL GET Request
- BeautifulSoup Object
- Pandas DataFrame

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Fa
lcon_Heavy_launches&oldid=1027686922"
```

First, let's perform an HTTP GET method to request the Falcon9 Launch HTML page, as an HTTP response.

```
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
```

Create a `BeautifulSoup` object from the HTML `response`

```
# Use BeautifulSoup() to create a BeautifulSoup object from a response text cont
ent
soup = BeautifulSoup(response.text)
```

Print the page title to verify if the `BeautifulSoup` object was created properly

```
# Use soup.title attribute
soup.title
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wikipedia</title>
```

https://github.com/AlvaroOrtiz2001/spacex-project/blob/main/jupyter-labs-webscraping%20(FINAL).ipynb

Data Wrangling

Perform exploratory data analysis eda on datatest

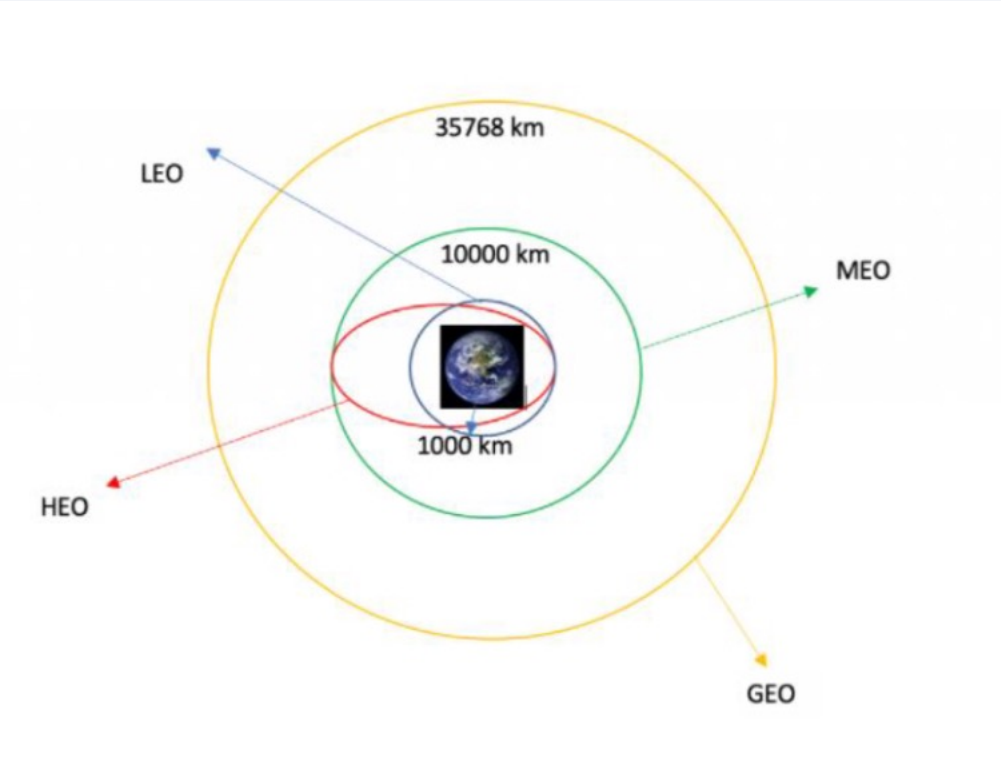
Calculate the number of launches at each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from outcome column

Work out success rate for every landing in dataset



https://github.com/AlvaroOrtiz2001/spacex-project/blob/main/labs-jupyter-spacex-Data%20wrangling%20(FINAL).ipynb

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:
- Mean vs Orbit

Line graph:
- Success rate vs Year

https://github.com/AlvaroOrtiz2001/spacex-project/blob/main/jupyter-labs-eda-dataviz%20(FINAL).ipynb

EDA with SQL

The SQL queries performed are:

- 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_outcomes in ground pad ,booster versions, launch_site for the months in year 2017

- Ranking the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

https://github.com/AlvaroOrtiz2001/spacex-project/blob/main/lab_jupyter_launch_site_location%20(FINAL).ipynb

Build an Interactive Map with Folium

- We took latitude and longitude coordinates of each launch site to visualize where it was each launch site

- We use some formulas in order to calculate the distance from launch sites to landmarks to find some trends

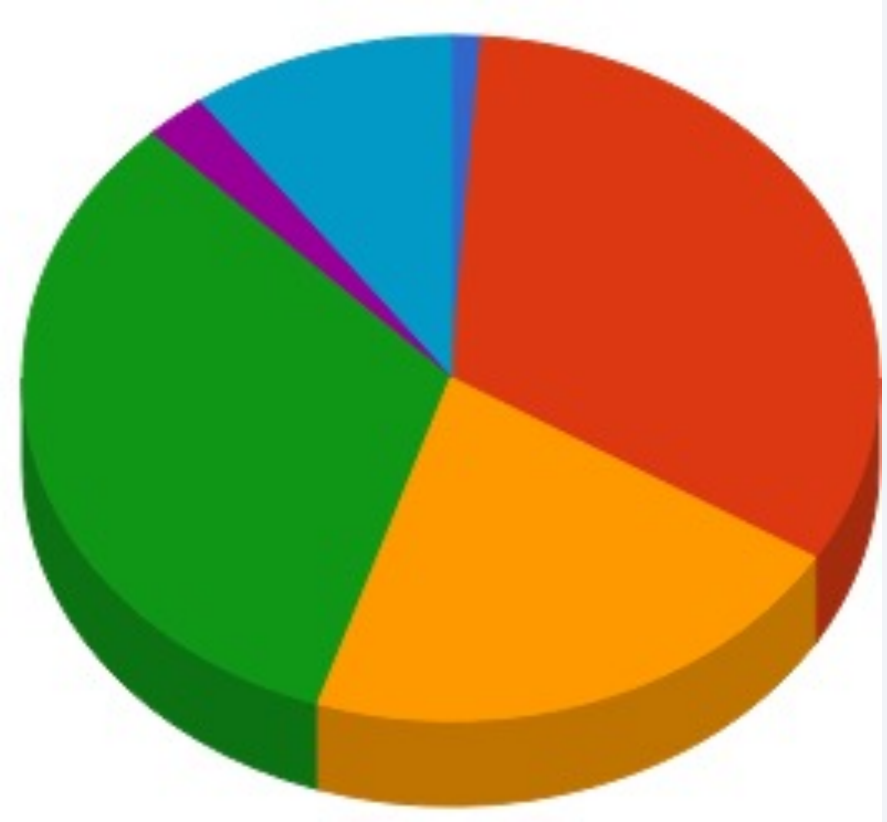
- Questions answered afeter building the interactive map

Are launch sites in close proximity to railways? No
Are launch sites in close proximity to highways? No
Are launch sites in close proximity to coastline? Yes
Do launch sites keep certain distance away from cities? Yes

https://github.com/AlvaroOrtiz2001/spacex-project/blob/main/lab_jupyter_launch_site_location%20(FINAL).ipynb

Build a Dashboard with Plotly Dash

Use of Pie Charts to show total launches by certain sites and percentages of the date



Predictive Analysis (Classification)

APDO MODEL

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

Fitour datasets into the GridSearchCV objects and train our dataset.

EVALUATING MODEL

Check accuracy for each model

Get tuned hyperparameters for each type of algorithms

Plot Confusion Matrix

IMPROVING MODEL

Feature Engineering

Algorithm Tuning

FINDING THE BEST PERFORMING CLASSIFICATION MODEL

The model with the best accuracy score wins the best performing model

Inthe notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

https://github.com/AlvaroOrtiz2001/spacex-project/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5%20(FINAL).ipynb

Results

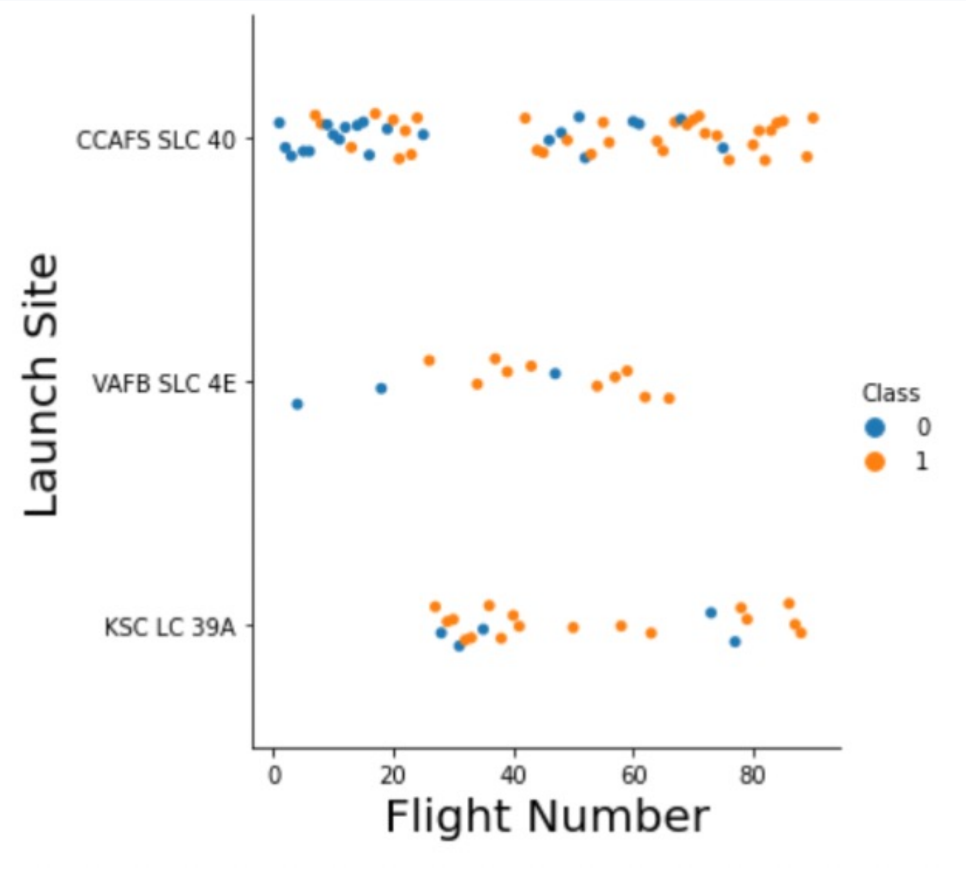
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

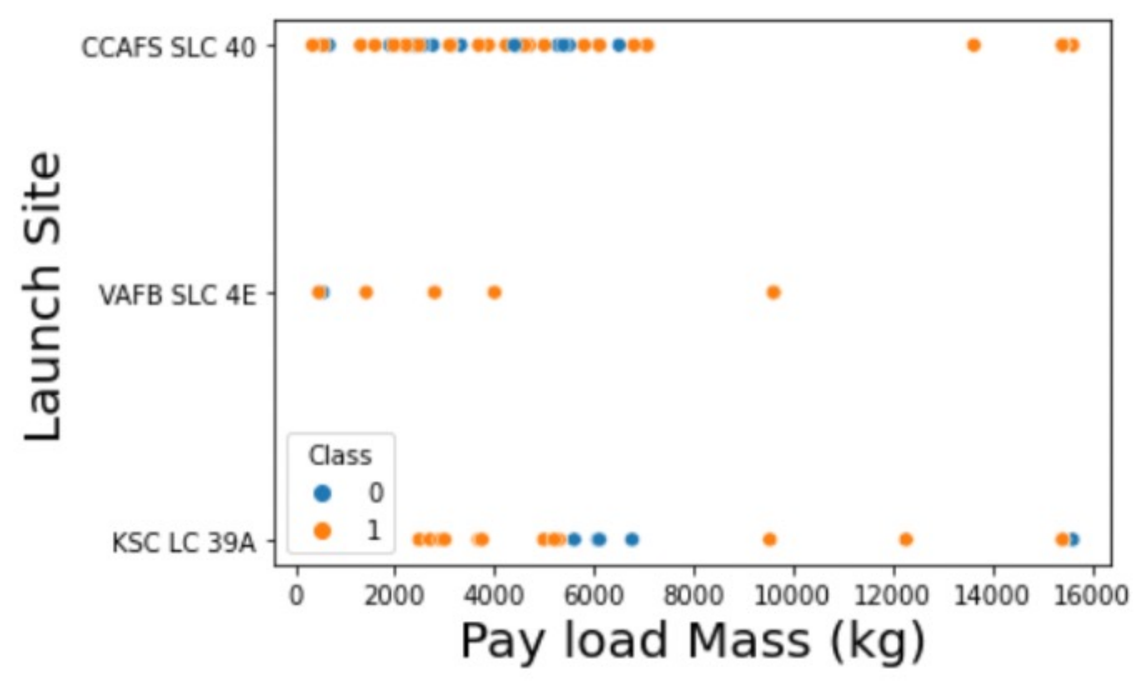
Flight Number vs. Launch Site

- CCAFS SCL 40 & a couple of VAFB SCL 4E were the launch sites of the first 20 flights
  - Most of them fail due they were the first attempts
- VAFB SCL 4E has a high success rate, however it is the launch site with the small amount of launches
  - The recent launches has in most of the cases successful
- This is because as they are more attempts, they improve the launches



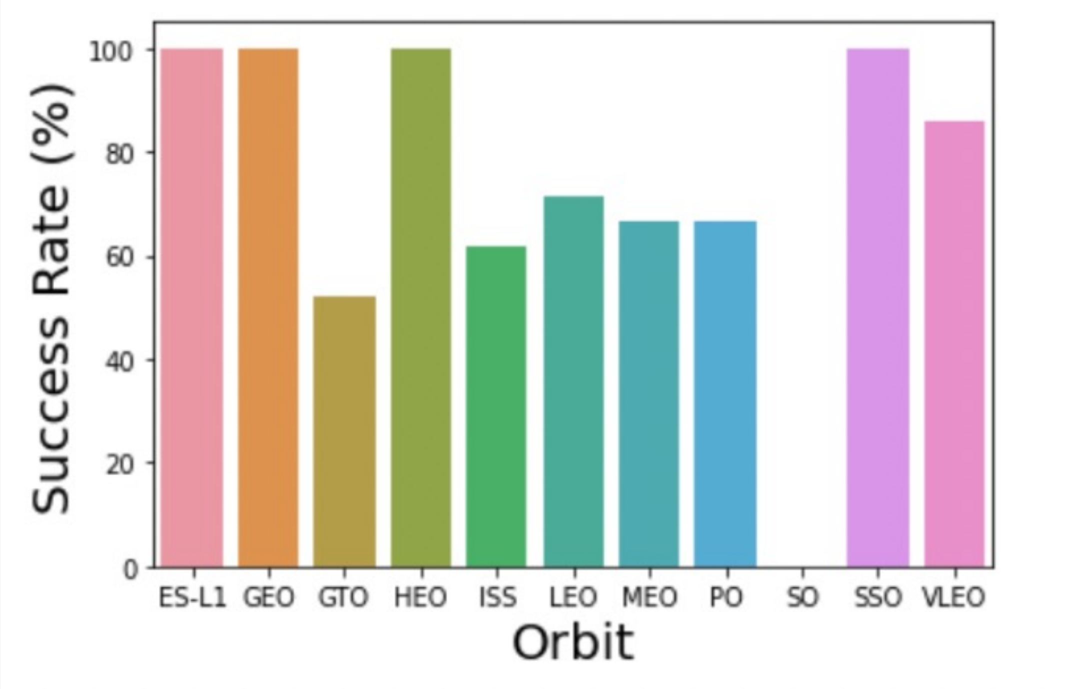
Payload vs. Launch Site

- Most of the launch sites have been with pay load mass of less tan 8000 kg
- Different cases depending of the launch site and the pay load mass
- CCAFS SLC 40
  - Results are not consistent, but the launches with higher pay load mass are more accurrate
- VAFB SLC 4E
  - The most consistent launch site, but not many launches
- KSC LC 39A
  - Majority of the failures in pay load mass around 6000kg so there should be a problem there



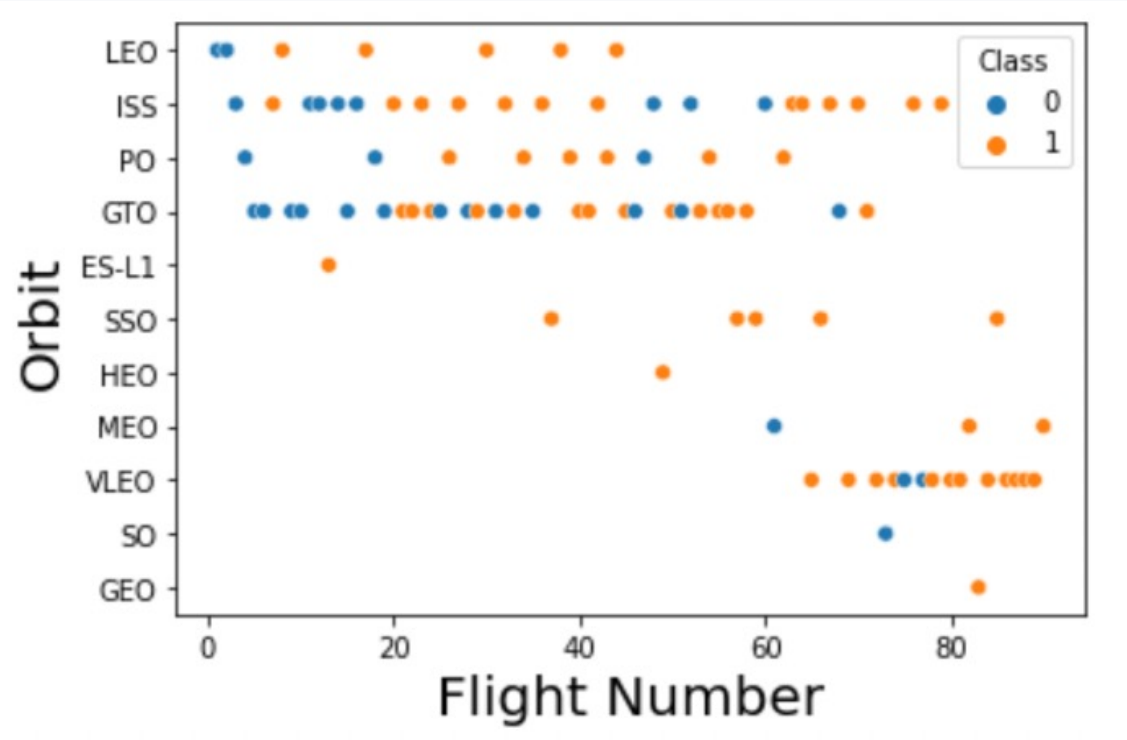
Success Rate vs. Orbit Type

- Different orbits have been selected and have been measured in % of success rate
    - 4 of them have a success rate of 100%, but there is an issue and is that we don't have the info of how many flights were to each orbit and it should be needed to determine the significance rate
    - 5 other orbits have a success rate of around 60%
    - An orbit called SO is the only orbit that has a success rate of 0%, meaning that it didn't succeed any landing



Flight Number vs. Orbit Type

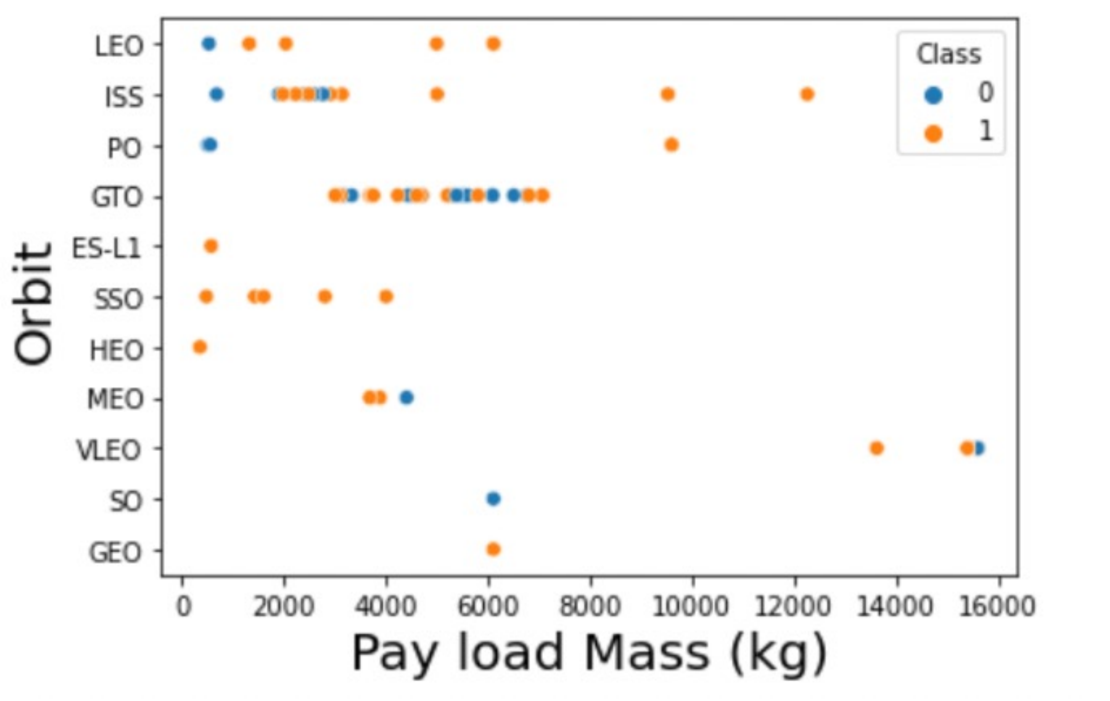
- First flights in: LEO, ISS, PO and GTO. Many of the failures occurred here
  - There is correlation between the first flights and failure
  - In LEO, ISS and PO, we see that after the first flights the successful landing percentage increase a lot
- VLOE Orbit, very consistent success. Could be correlated to the fact that are last flights



Payload vs. Orbit Type

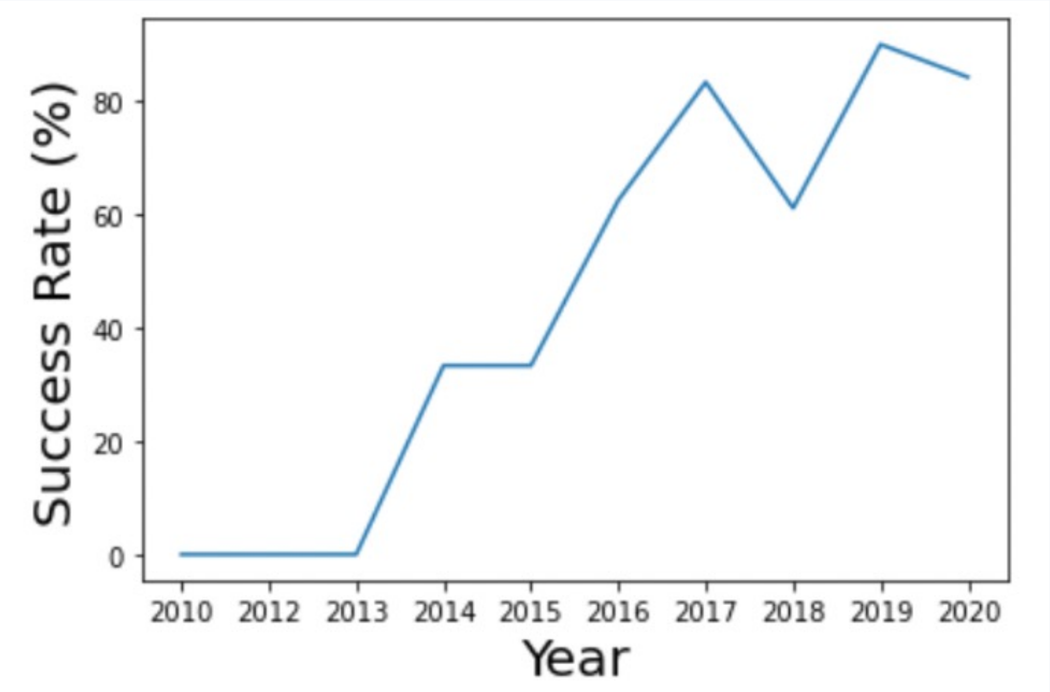
- As we show in the pad load mass and launch site correlation, the highest pay load mass are more consistent with the success rate

- Flight in SSO all with different masses an all of them with success rate



Launch Success Yearly Trend

- There exists a positive linear correlation between year of launch and success rate. As year passes, the success rate increase
- Something happen in 2018, because the success rate drop from 80% to 60% to then keep increasing the success rate
  - Factors of this could be external conditions, exploration of new orbits or new use of different payload masses



All Launch Site Names

There are only 4 launch sites for the 90 launches

SELECT DISTINCT launch_site

FROM SPACEXTBL

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

Launch Site Names Begin with 'CCA'

The first 5 launches of CCAFS LC-40 were successful

SELECT*

FROM SPACEXTBL

WHERE launch_site LIKE 'CCA%'

LIMIT 5

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

Total Payload Mass

The payload mass of NASA (CRS) boosters are much heavier than other

```
SELECT SUM(payload_mass:kg_) AS Total_Payload

FROM SPACEXTBL

WHERE customer = ‘NASA (CRS)’
```

total_payload
45596

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2928.4 kg, this implies us that heavier payload mass is better

```
SELECT AVG(payload_mass_kg_) AS Avg_payload

FROM SPACEXTBL

WHERE booster_version = 'F9 v1.1'
```

avg_payload
2928.400000

First Successful Ground Landing Date

As flights started in 2010, this means that it took almost 6 years until the first rocket succeed

```
SELECT MIN(DATE) AS First_Success

FROM SPACEXTBL

WHERE landing__outcome = ‘Success (ground pad)’
```

first_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

There have been only 4 successful trials between this range, this supports the argument that have been said in previous slides that there has been an issue with flights with a payload mass around 6000 kg

payload	payload_mass__kg_	landing__outcome
JCSAT-14	4696	Success (drone ship)
JCSAT-16	4600	Success (drone ship)
SES-10	5300	Success (drone ship)
SES-11 / EchoStar 105	5200	Success (drone ship)

SELECT payload, payload_mass__kg_,

Landing outcome

FROM SPACEXTBL

WHERE lading__outcome = ‘Success (drone ship)’

AND payload_mass_kg_ BETWEEN 4000 and 6000

Total Number of Successful and Failure Mission Outcomes

We have seen that there has been many failures in the flights, however, there has been success in most of the objectives of those flights

```
SELECT mission_outcome, COUNT(*) AS Total

FROM SPACEXTBL

GROUP BY mission_outcome
```

mission_outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The only booster version that carried the maximum payload mass has been the F9 booster.

```
SELECT booster_version, payload_mass___kg_
FROM SPACEXTBL
WHERE payload_mass__kg =
        (SELECT MAX(payload_mass__kg_)
        FROM SPACEXTBL)
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

Two attempts ended with failed drone ship landings.

SELECT date, booster_version, launch_site, landing__outcome

FROM SPACECTBL

WHERE landing__outcome = ‘Failure (drone ship)’

AND YEAR(DATE) = 2015

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

It is interesting that there was not even attempt to landing in 10 of this trials, therefore this trials should not count in the landing data

SELECT landing__outcome,

COUNT(landing__outcome) AS Total

FROM SPACEXTBL

WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'

GROUP BY landing__outcome

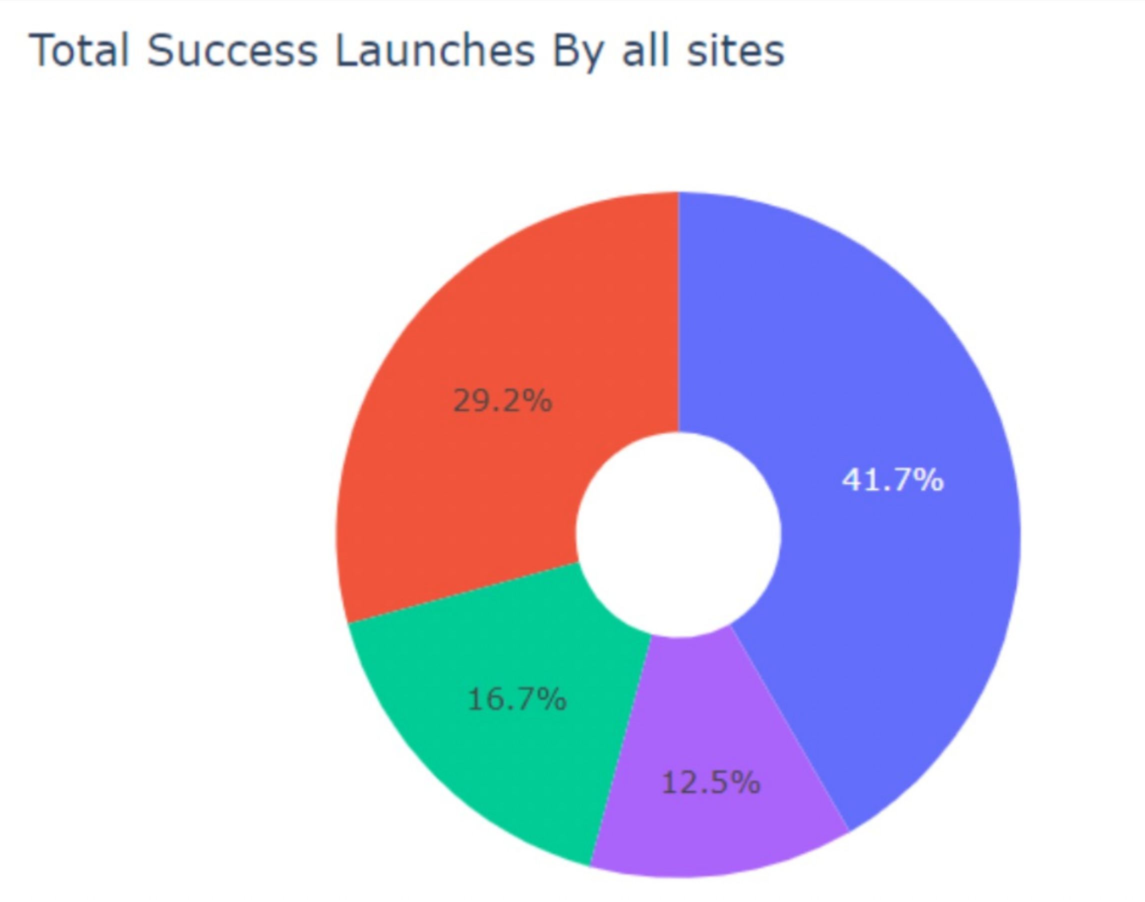
ORDER BY Total DESC

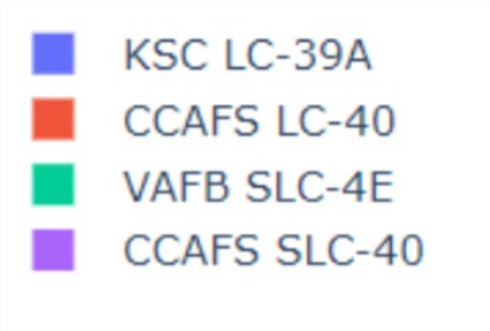
landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 5

Build a Dashboard with Plotly Dash

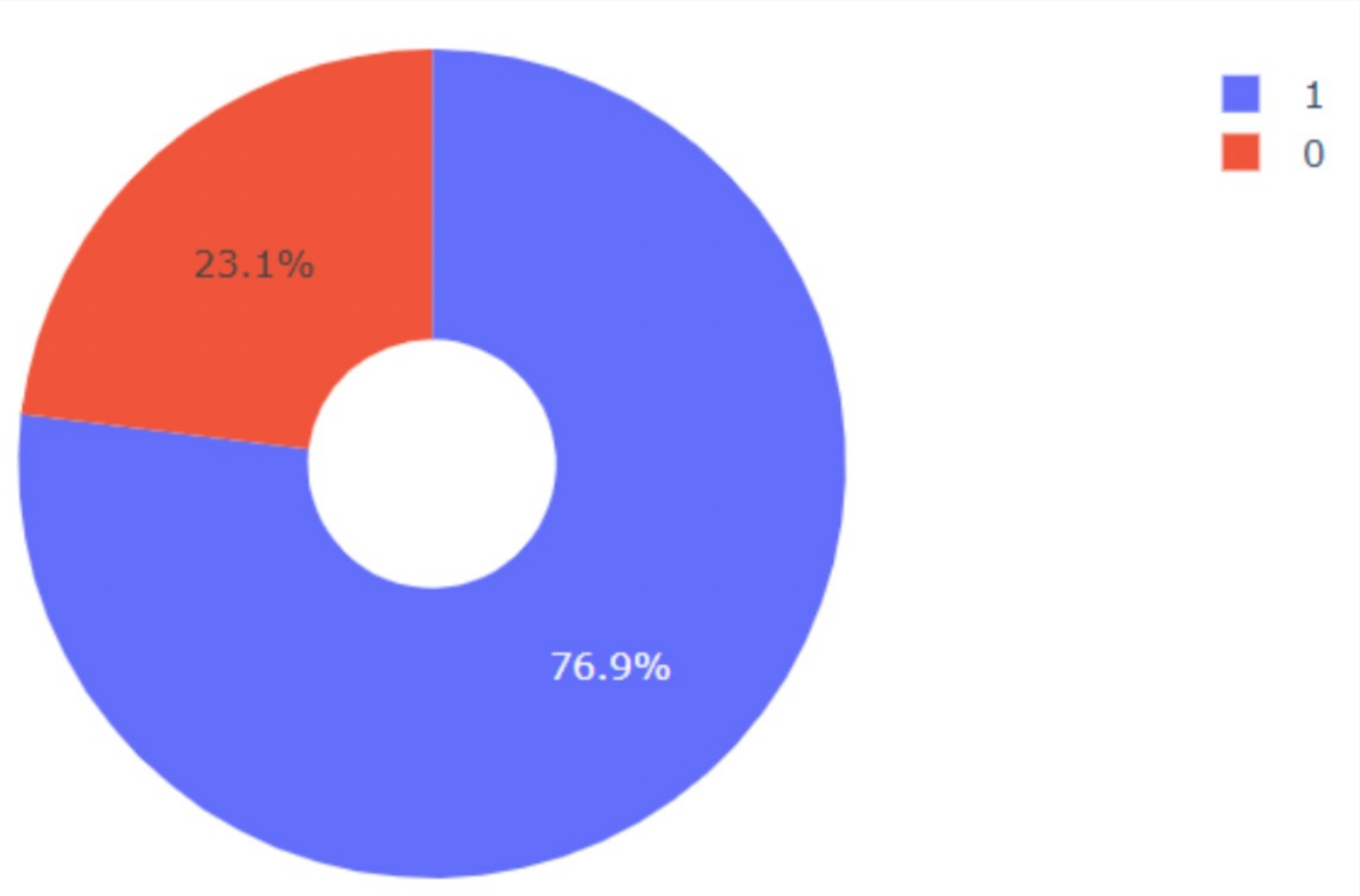
Dashboard – Pie chart % success rate by launch site





KSC LC-39A is the most successful one from all the sites

Dashboard – Pie chart of the launch site with highest launch success ratio



In this case:

- 76.9% success
- 23.1% failure

Section 6

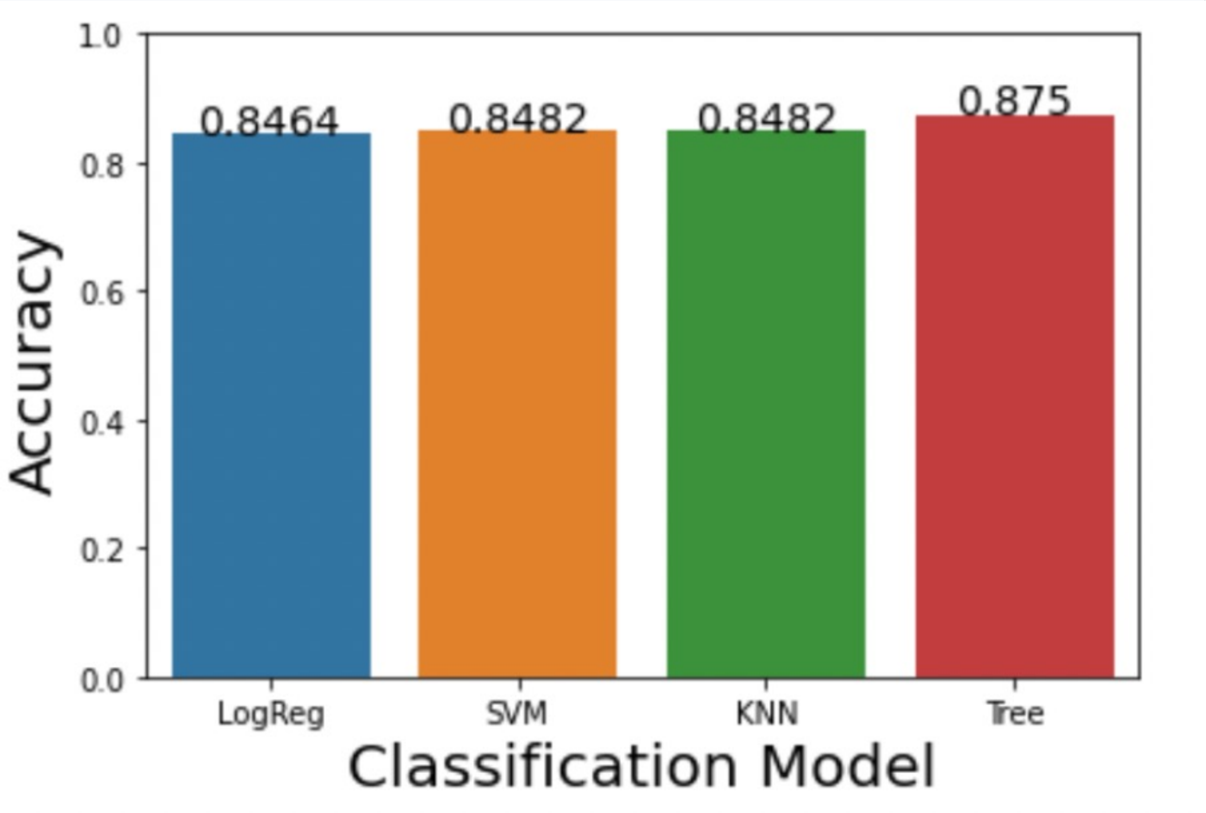
Predictive Analysis (Classification)

Classification Accuracy

4 classification models were designed with different steps:

- Standardization of the feature dataset
- Hyperparameter tuning with GridSearchCV
- Model fitting with training dataset

The most accurate model according to the graph that one see on the right is the Tree model



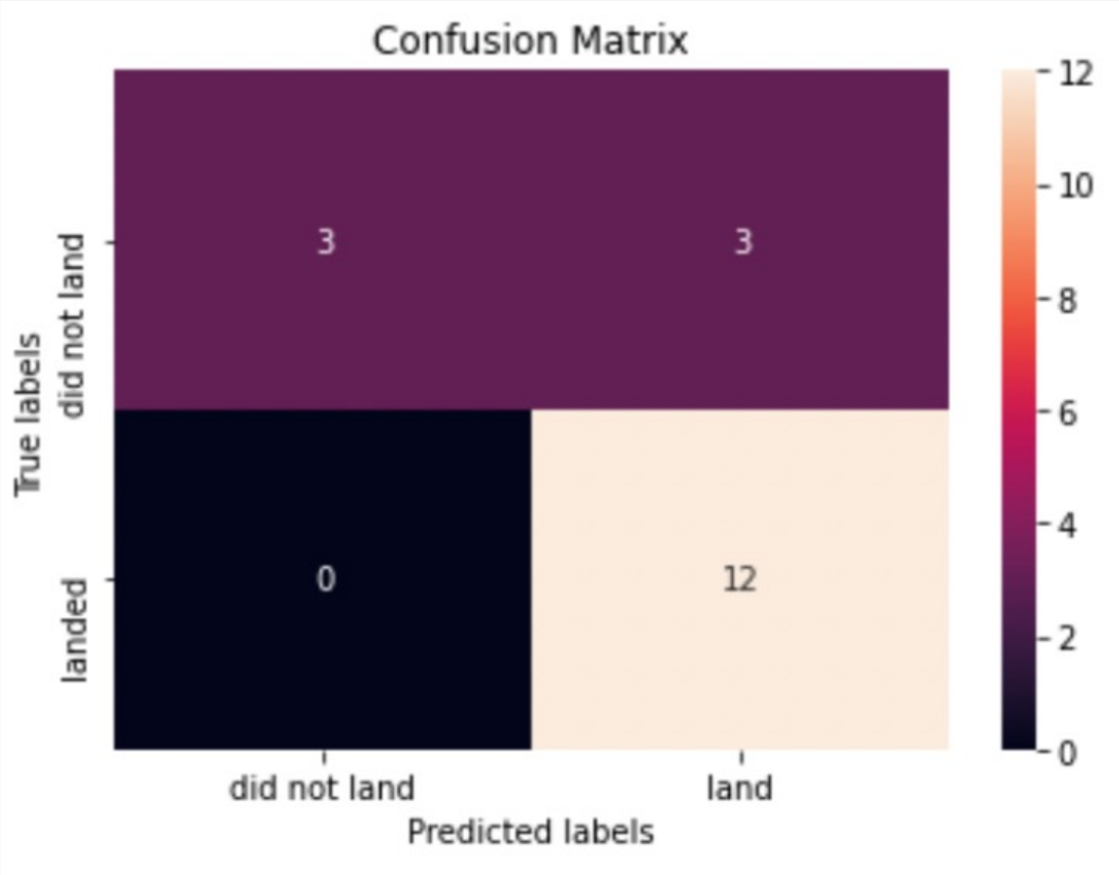
Confusion Matrix

Now one is going to see the confusion matrix and the information that we can take of the more accurate model, in this case the Decision Tree model.

- Accuracy = 0.83

- Recall (Type I Error) = 1.0

- Precision (Type II Error) = 0.8



Conclusions

- It took many years until we saw the first successful landing of SPACEX, however data science and machine learning can help us to determine what missions are going to fail and it will save us time and money
- In this project, external factors such as weather conditions or different technical components are excluded, and there should be needed in order to create a more accurate model
- Finally, with the rate of success that we have seen during the recent trials we are close to see commercial space in the next few years

Thank you!