IBM Developer
SKILLS NETWORK

# Winning Space Race with Data Science

Ferradas Martin
11/9/2022

# Outline

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

# Executive Summary

- **Summary of methodologies**
  - *Data Collection:* Web scraping and information gathering using the SpaceX API
  - *Data analysis:* Data wrangling and data visualization (including interactive analytics)
  - *Outcome prediction:* Machine learning using different algorithms.
- **Summary of all results**
  - *Data was retrieved from the public sites listed above, processed and visualized. A machine learning algorithm was deployed to predict the success of launchings based on multiple variables achieving an accuracy of 94%*

# Introduction

- **Project background and context**
  - So far Space X has been a very successful company given the low cost of its operations compared to the competition.
  - A Big part of that cost difference is due to the fact that part of the ship can be recovered after being launched.  If the success rate increases, the cost will decrease too.
  - We want to perform a study to understand how to maximize the success rate of the operation to be used on the establishment of a new corporation space Y.
- **Desired outcomes:**
  - Evaluate the possibility to estimate if a launch would be successful or not as the success rate has a direct impact on the cost of operations of the company.
  - Determine what is the best place to make launches.

Section 1

# Methodology

# Methodology

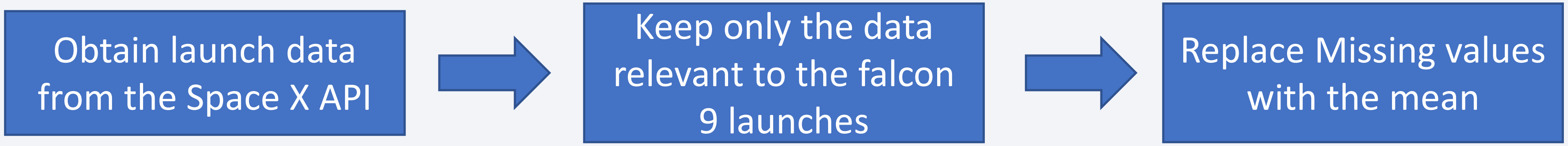
## Executive Summary

- Data collection methodology:
  - Sources:
    - Space X API: https://api.spacexdata.com/v4/
    - WebScraping: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - Enriched the data by creating a landing outcome column based on outcome data
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotty Dash

# Data Collection

- Perform predictive analysis using classification models
  - After collecting the data, it was normalized and converted to numeric format. The data available was divided in a training and testing set. Multiple classification models were trained and tested with both sets using the grid search algorithm to find the best hyper-parameters for the individual algorithm. After that the algorithm with the best accuracy was selected as the one to use on the future.
- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

# Data Collection – SpaceX API




**GitHub Repository:**

https://github.com/mferrad/coursera/blob/main/python/Capstone/jupyter-labs-spacex-data-collection-api.ipynb

# Data Collection - Scraping

Request the Falcon9 Launch Wiki page from its URL → Extract all column/variable names from the HTML table header → Create a data frame by parsing the launch HTML tables

**GitHub Repository:**

https://github.com/mferrad/coursera/blob/main/python/Capstone/jupyter-labs-webscraping.ipynb

# Data Wrangling

- Firstly a data analysis was performed on the dataset to understand possible orbits, outcomes, etc.
- After that, the information was summarized to understand how often the different possible values repeat.
- Finally, the landing outcome label was created based on the original outcome feature.

**GitHub Repository:**

https://github.com/mferrad/coursera/blob/main/python/Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

# EDA with Data Visualization

- The following plots were used to understand the relationship between variables and look for correlations between the features. One of the important things that were analyzed was how the orbit affected the success rate and if the success rate was improving per year.
  - **Scatter plots:** FlightNumber vs. PayloadMass, FlightNumber vs LaunchSite, PayloadMass vs LaunchSite, FlightNumber vs. Orbit, PayloadMass vs Orbit
  - **Bar charts:** Success rate per Orbit
  - **Line chart:** Success rate per year

**GitHub Repository:**

https://github.com/mferrad/coursera/blob/main/python/Capstone/jupyter-labs-eda-dataviz.ipynb

# EDA with SQL

## The following SQL queries were performed:

- Names of the unique launch sites in the space mission
- Top 5 launch sites whose name begin with the string 'CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date when the first successful landing outcome in ground pad was achieved
- Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
- Total number of successful and failure mission outcomes
- Names of the booster versions which have carried the maximum payload mass
- Failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank of the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20.

**GitHub Repository:**

https://github.com/mferrad/coursera/blob/main/python/Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

# Build an Interactive Map with Folium

- Markers, circles, lines were used with folium maps:
  - Markers were used to indicate key points on the map such as lunch stations
  - Circles were used to highlight specific areas
  - Lines were used to indicate the closest road between 2 points.

**GitHub Repository:**

https://github.com/mferrad/coursera/blob/main/python/Capstone/lab_jupyter_launch_site_location.ipynb

# Build a Dashboard with Plotty Dash

- Two graphs and plots were put together
  - Percentage of launches per site
  - Payload range
- Both graphs were used to quickly visualize how the launces and the payload affected the success rate and how often the combination happened.

**GitHub Repository:**

https://github.com/mferrad/coursera/tree/main/python/Capstone/Dashboard_with_plotty_dash

# Predictive Analysis (Classification)

- Four Models were compared in order to get the best predictive model: Logistic regression, support vector machine, decision tree and K nearest neighbors.

Data preparation + split into test and training → Grid search to find best hyperparameters + model testing → Results comparison among the models

**GitHub Repository:**

https://github.com/mferrad/coursera/blob/main/python/Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

https://github.com/mferrad/coursera/blob/main/python/Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5_NO_OUTPUT.ipynb

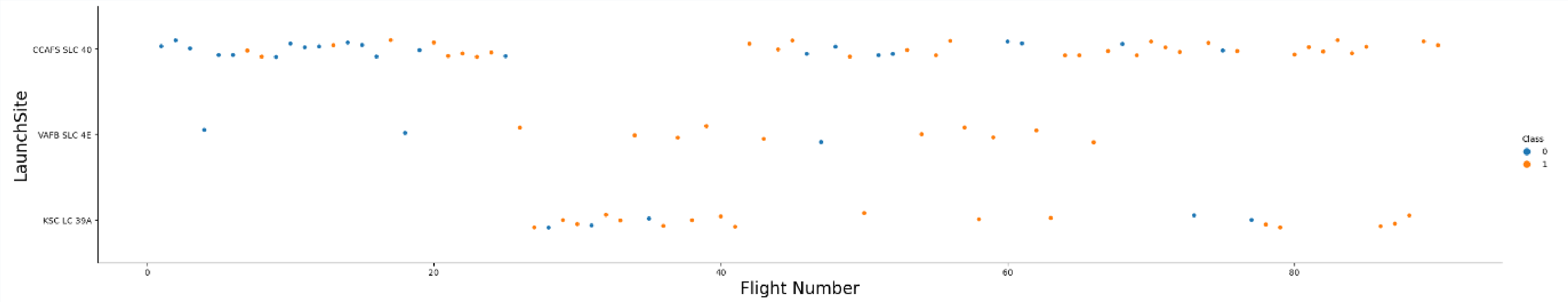
# Results

- First success landing outcome happened in 2015
- Successful landing rate becomes higher every year
- Many Falcon 9 booster versions were successful even with payload above average
- The orbits ES-L1, GEO, HEO and SSO have a perfect success rate so far
- KSC LC-39 was the site with the highest number of successes and a success rate of almost 77%
- The decision tree model performed the best obtaining an accuracy of almost 95%

Section 2

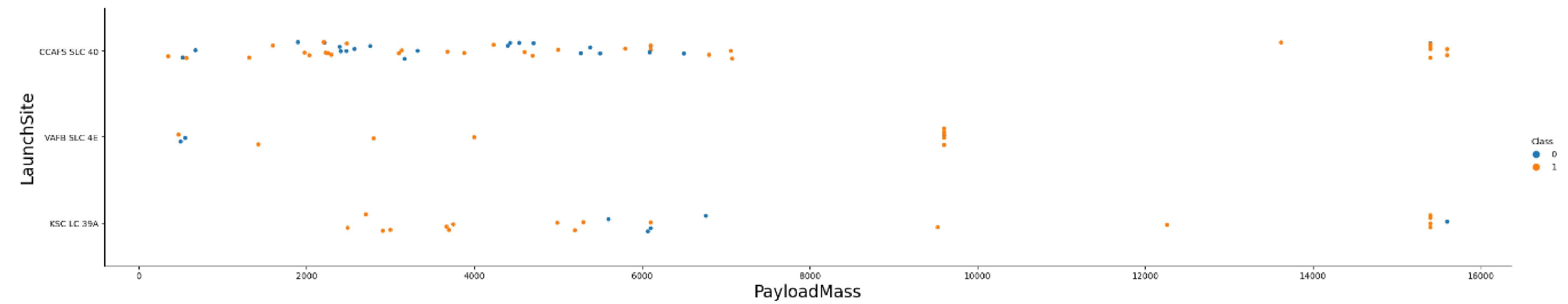
# Insights drawn from EDA

# Flight Number vs. Launch Site




- We can see that the lunch sites have improved their success rate over time.
- CCAF5 SLC 40 is the one that had the most consecutive successful launches recently.

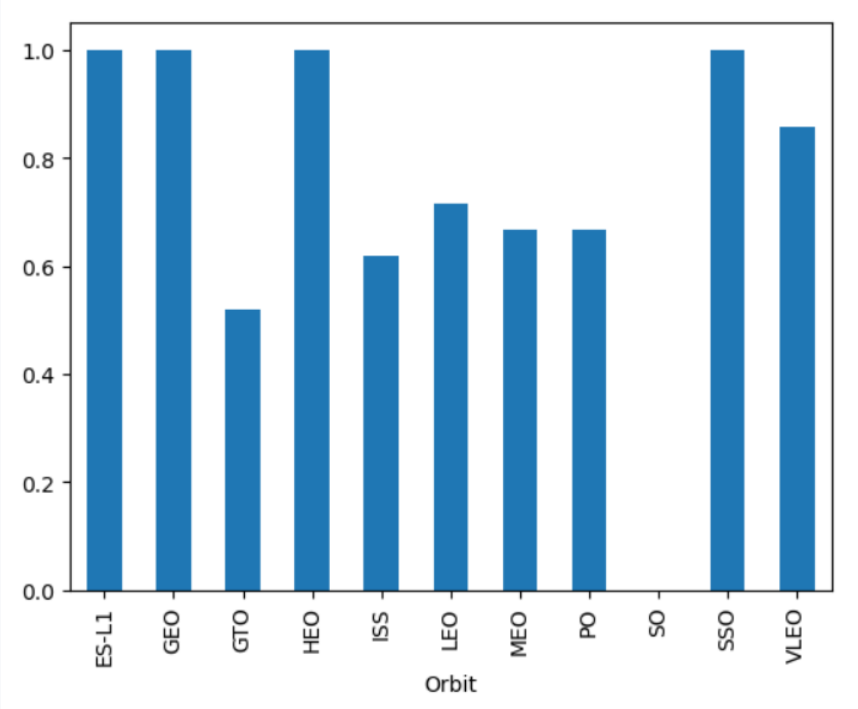
# Payload vs. Launch Site



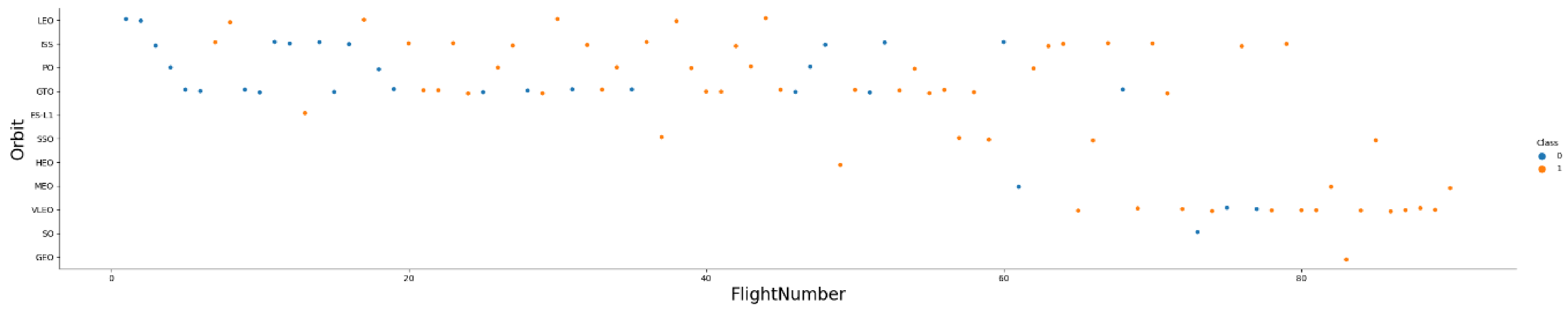

- As the payload increases the success rate increases as well. Almost all payloads over 8000kg were successful.
- VAFB SLC 4E has not managed payloads over 1000kg, however, it is almost always successful for the different ranges of payloads it does cover.

# Success Rate vs. Orbit Type

- The orbits ES-L1, GEO, HEO and SSO have a perfect success rate so far.
- There has not been a launch that had SO as a target.
- All the other orbits have different ranges of success rates.

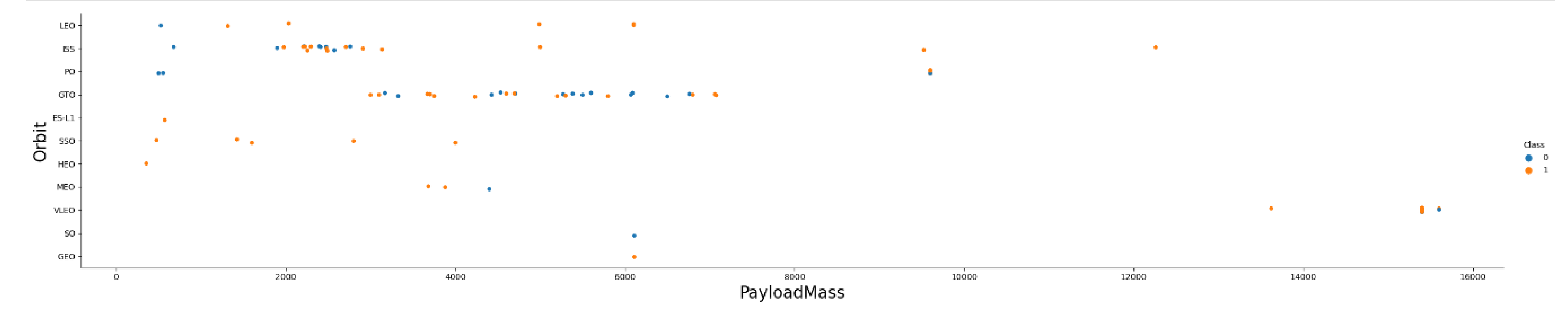


# Flight Number vs. Orbit Type




- Success rate has been increasing for all orbits.
- Latest launches have been targeting the VLEO orbit with high success rate.

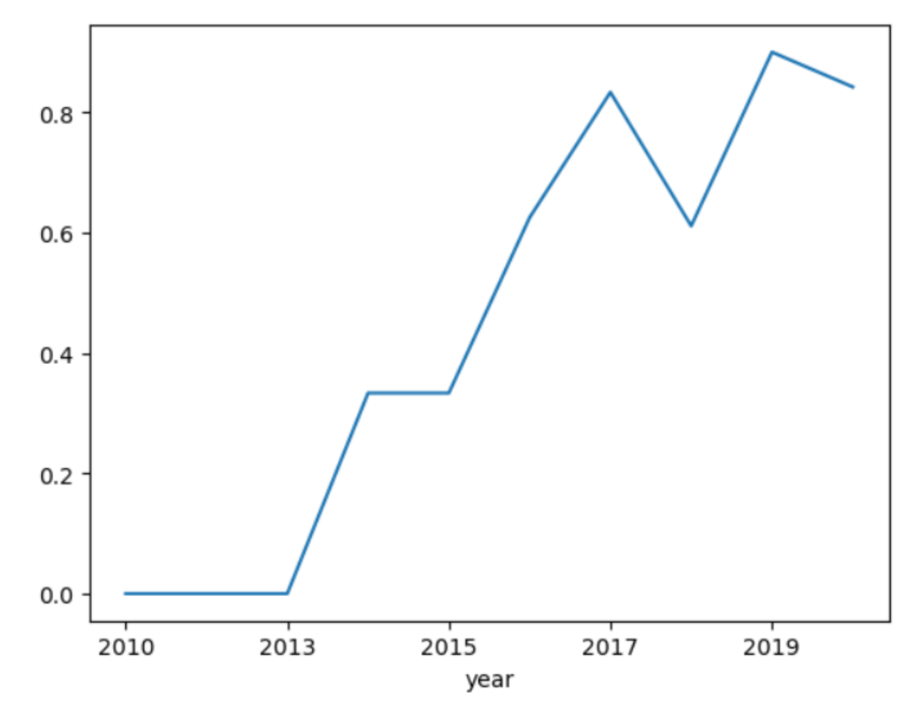
# Payload vs. Orbit Type




- Different orbits do not cover a unique range of payload mass.
- For most of the orbits a specific payload mass and a specific orbit does not seem to guaranty the success.

# Launch Success Yearly Trend

- Success rate has been increasing with time.
- 2018 is the only year that has not shown an increase of success rate.



# All Launch Site Names

- Lunch sites:

| Launch_Sites |
|---|
| CCAFS LC-40 |
| VAFB SLC-4E |
| KSC LC-39A |
| CCAFS SLC-40 |

- There are 4 different launch sites on the dataset.

# Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`

| Date | Time (UTC) | Booster_Version | Launch_Site | Payload | PAYLOAD MASS kg | Orbit | Customer | Mission Outcome | Landing _Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 04-06-2010 | 18:45:00 | F9 v1.0 B0003 | CCAFS LC-40 | Dragon Spacecraft Qualification Unit | 0 | LEO | SpaceX | Success | Failure (parachute) |
| 08-12-2010 | 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) |
| 22-05-2012 | 07:44:00 | F9 v1.0 B0005 | CCAFS LC-40 | Dragon demo flight C2 | 525 | LEO (ISS) | NASA (COTS) | Success | No attempt |
| 08-10-2012 | 00:35:00 | F9 v1.0 B0006 | CCAFS LC-40 | SpaceX CRS-1 | 500 | LEO (ISS) | NASA (CRS) | Success | No attempt |
| 01-03-2013 | 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 total payload maps corresponding to the customers ‘NASA (CRS)’ was 45596 kg

| sum(PAYLOAD_MASS__KG_) |
|---|
| 45596 |

# Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 V1.1 is 2928.4 kg

| avg(PAYLOAD_MASS__KG_) |
| --- |
| 2928.4 |

# First Successful Ground Landing Date

- First successful landing outcome on ground pad:

| min(Date) |
|---|
| 01-05-2017 |

- Filtering data by successful landing on ground pad and getting the minimum value for date you get the result above.

# Successful Drone Ship Landing with Payload between 4000 and 6000

- Boosters which 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 |

- The result above was obtained after filtering the landing condition and payload requirements. There are only 4 booster versions matching the criteria.

# Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:

| Mission_Outcome | Number |
|---|---|
| Failure (in flight) | 1 |
| Success | 99 |
| Success (payload status unclear) | 1 |

- The results above are obtained by grouping the records by outcome and adding the number of records per group. We can see that the success rate is almost 100%

# Boosters Carried Maximum Payload

- The table on the right shows the boosters that carried the maximum payload

- The result is obtained by calculating the maximum payload for the entire dataset and matching the booster versions that had that payload at some point.

- 12 booster versions have carried the maximum payload.

| 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

- List of the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

- There are only 2 launch records for 2015

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

- The table on the right shows 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

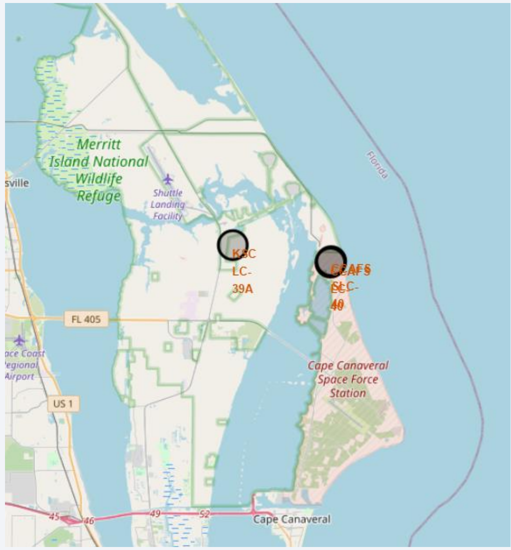
- We can see that the different “flavors” of success are ranked higher than the failures.

| Landing _Outcome | rank |
|---|---|
| Success | 20 |
| No attempt | 10 |
| Success (drone ship) | 8 |
| Success (ground pad) | 6 |
| Failure (drone ship) | 4 |
| Failure | 3 |
| Controlled (ocean) | 3 |
| Failure (parachute) | 2 |
| No attempt | 1 |

Section 3

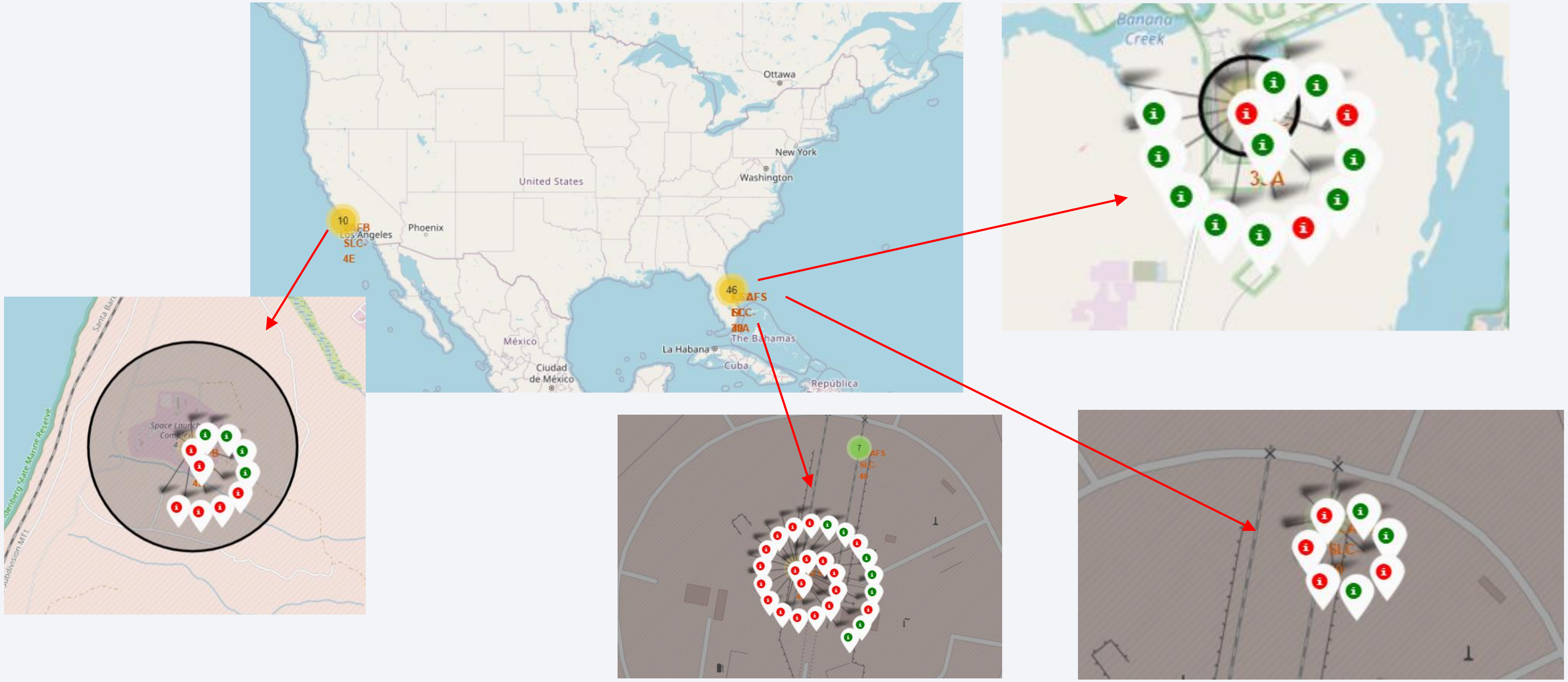
# Launch Sites Proximities Analysis

# Launch sites

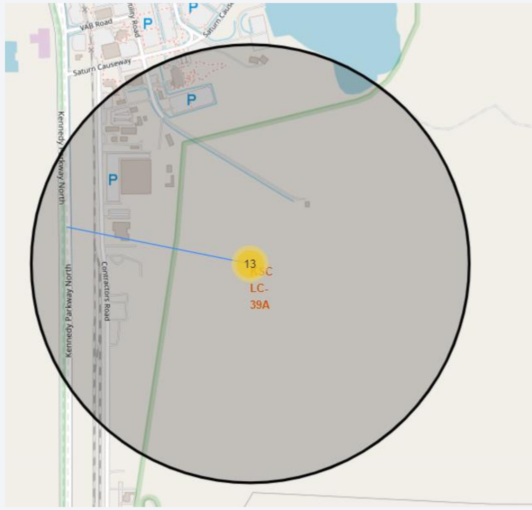

- Launch sites are located close to the coast for security reasons.

# Launch outcome per site




- Successful launches in green, failed launches in red

# <Distance to sea / highway >

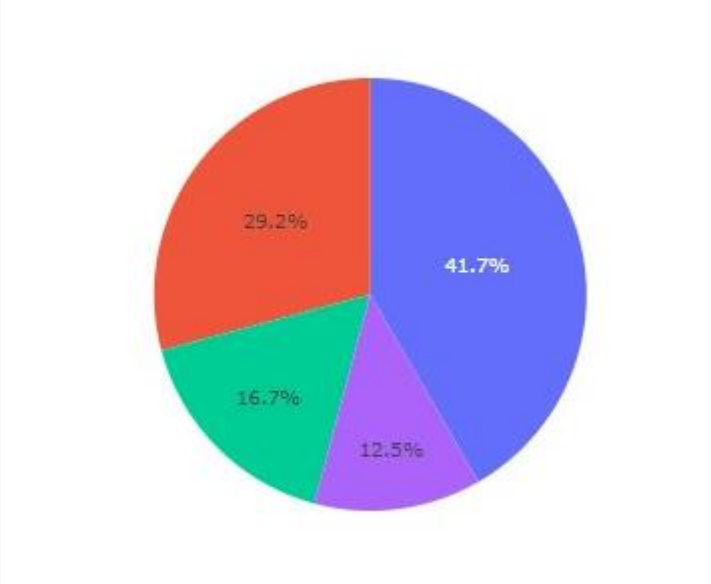



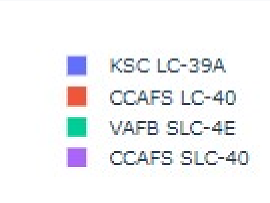

- Sites are close to the sea and near roads but far away from cities.

Section 4

# Build a Dashboard with Plotly Dash

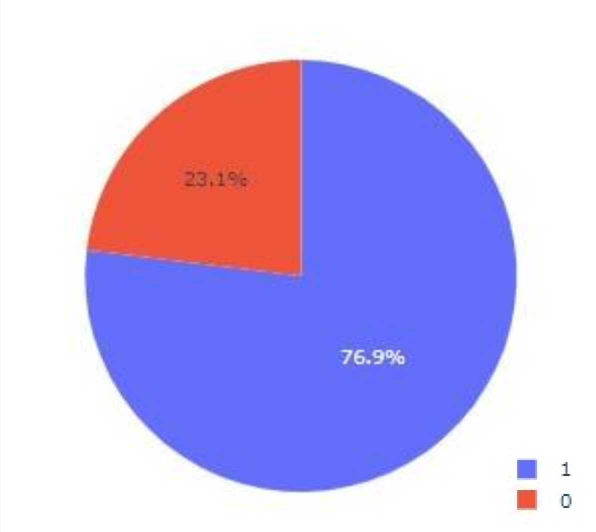
# Launch success per site



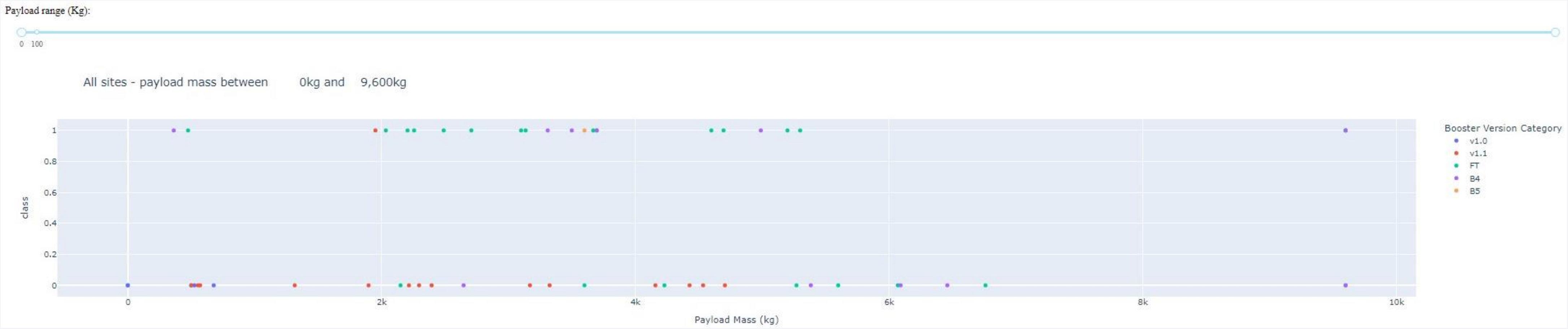

- We can see that KSC LC-39 was the site with the highest number of successes

# Success rate for KSC LC-39A




- At the time of the analysis KSC LC-39A has almost 77% success rate

# Payload vs. Launch Outcome for all sites




- Payloads under 6000kgs and FT boosters have the highest success rate

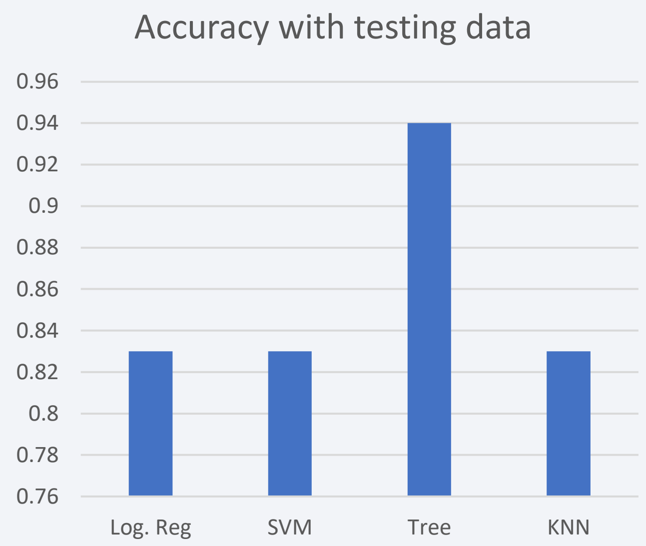
Section 5

# Predictive Analysis (Classification)

# Classification Accuracy

- The figure on the right shows the accuracy of the different models tested with the testing dataset after they were optimized with the training dataset.

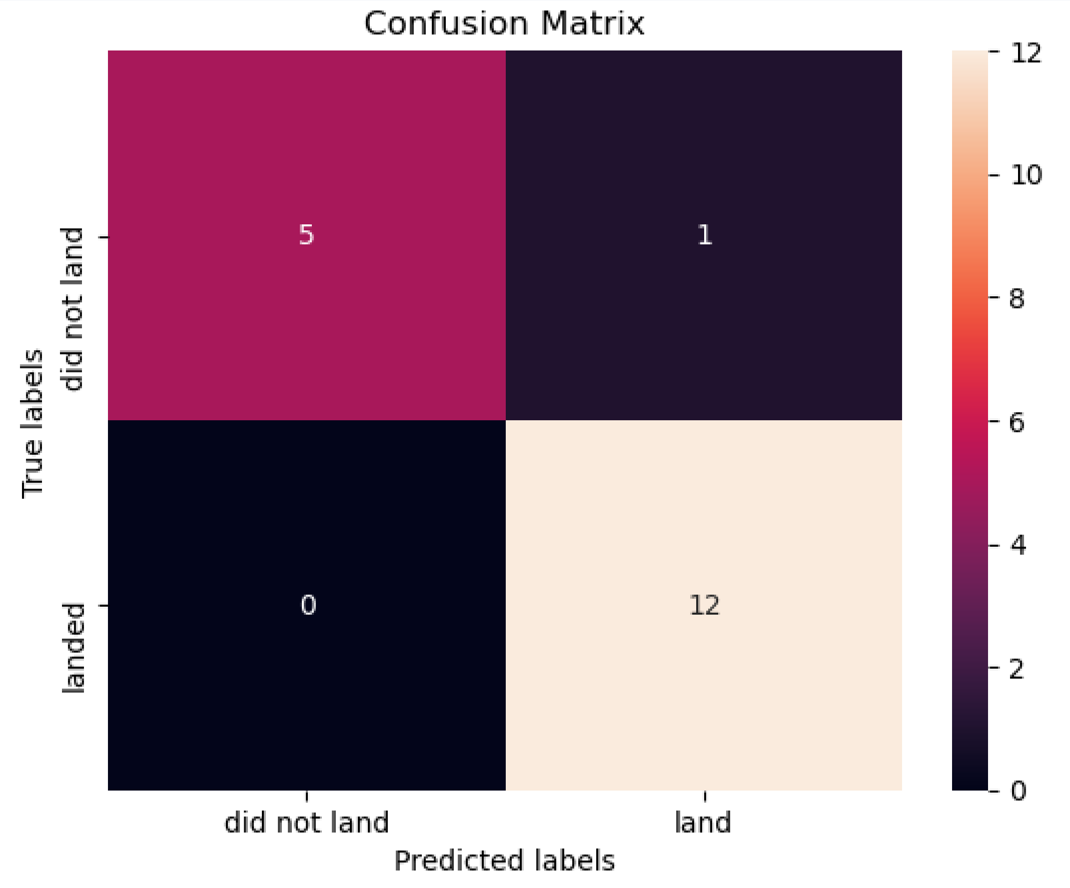
- The decision tree classifier shows the highest accuracy with almost 95%



# Confusion Matrix for the decision tree classifier

- The figure on the right shows the confusion matrix for the decision tree classifier.

- We can see that the estimation is almost perfect (almost all the values land in the diagonal) except for one incorrect landing prediction.



# Conclusions

- An exhaustive analysis was performed using the information gathered from the Space X API and the launches information listed on Wikipedia.
- Successful landing rate becomes higher every year.
- The orbits ES-L1, GEO, HEO and SSO have a perfect success rate so far.
- KSC LC-39 was the site with the highest number of successes and a success rate of almost 77%.
- The decision tree model performed the best obtaining an accuracy of almost 95%.

# Appendix

- Repository for the entire project:
  https://github.com/mferrad/coursera/tree/main/python/Capstone

# Thank you!