Applied Data Science Capstone

SpaceX Landing Success Prediction

Name: Marwane Hmidi

Date: 26 December 2022

OUTLINE

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

EXECUTIVE SUMMARY

- In this capstone project, we will make a prediction on whether the SpaceX Falcon 9 first stage will land. We can calculate the cost of a launch once we know whether the first stage will land. Different machine learning classification techniques will be used to achieve this.
- Data collection, data wrangling and preprocessing, exploratory data analysis, data visualization, and machine learning prediction will all be part of the methodology used.
- The findings of our inquiry and analysis suggest that there are specific characteristics of rocket launches that are correlated with successful or unsuccessful launches.
- In the end we conclude that the Decision Tree may be the best machine learning algorithm to for this problem.

INTRODUCTION

- This capstone project's major objective is to predict whether the Falcon 9 first stage will successfully land. SpaceX advertises on their website that their rocket launches cost 62 million while other providers charge upwards of 165 million because they take great pride in being able to reuse the first stage of a rocket launch. The reuse of the first stage is largely responsible for these cost savings. The price of a launch can be calculated if we can know if the first stage will land. If a different business want to compete with SpaceX for a rocket launch, it may use the information provided here.
- This gets us to the key inquiry we are attempting to address: Under the specified conditions of a Falcon 9 rocket launch, will the rocket's first stage successfully land?

METHODOLOGY

- Data was gathered using two techniques: webscraping launch information from a Wikipedia article and requesting information from the SpaceX API. The data was then transformed and cleaned using the Pandas module in Python.
- Exploratory data analysis (EDA) was performed on the clean data utilizing visualization tools including Python's matplotlib and seaborn packages, as well as SQL queries to provide answers. In order to respond to some analytical queries, interactive visualization packages in Python were employed. Maps were produced using Folium, and interactive data visualizations with Plotly Dash.
- For the prediction study, four alternative machine learning classification models were employed. The models utilized were decision tree classifier, logistic regression, support vector machines, and k-nearest neighbor. To choose the best model, each one was trained, adjusted, and tested.

Data Collection – SpaceX API

1. Request and parse the SpaceX launch data using the GET request.
2. Normalize JSON response into a dataframe.
3. Extract only useful columns using auxilary functions.
4. Create new pandas dataframe from dictionary.
5. Filter dataframe to only include Falcon 9 launches.
6. Handle missing values.
7. Export to CSV file.

Github URL: Data Collection API

Data Collection - Scraping

1. Request rocket launch data from its Wikipedia page.
2. Extract all column/variable names from the HTML table header.
3. Create a data frame by parsing the launch HTML tables.
4. Export to CSV file.

Github URL: Webscraping

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

Github URL: Data Wrangling

EDA with Data Visualization

- Scatter plots: Scatter plots were used to represent the relationship between two variables. Different sets of features were compared such as Flight Numbers vs. Launch Site, Payload vs. Launch Site, Flight Number vs. Orbit Type and Payload vs. Orbit Type.
- Bar chart: Bar charts were used makes it easy to compare values between multiple groups at a glance. The x-axis represents a category and the y-axis represents a discrete value. Bar charts were used to compare the Success rate for different Orbit Types.
- Line chart: Line charts are useful for showing data trends over time. A line chart was used to show Success Rate over a certain number of years.

Github URL : EDA 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 'CCA' , 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 when the first successful landing outcome in ground pad was achieved, listing the names of the boosters which have success in drone ship 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 failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015, ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Github URL: EDA with SQL

Interactive Map with Folium

This parts consists of the creation and addition of objects to a Folium map. All launch sites, as well as the successful and unsuccessful launches for each site, were displayed on a map using marker objects. The distances between a launch location and its surroundings were calculated using line objects.

By adding these objects, the following geographical patterns about launch sites were found:

- Launch sites in close proximity to railways.
- Launch sites in close proximity to highways.
- Launch sites in close proximity to coastlines.
- Launch sites keep certain distance away from cities and inhabited areas.

Github URL: Interactive Map with Folium

Dashboard with Plotly Dash

- A pie chart displaying each site's successful launch. This graph is helpful since it allows you to display the success rate of launches on specific sites or visualize the distribution of landing outcomes across all launch locations.
- A scatter diagram illustrating the relationship between landing success and the mass of various boosters. The site(s) and payload mass are the dashboard's two inputs. This chart is helpful since it allows you to see how different factors influence the results of the landing.

Github URL: SpaceX Dashboard

Predictive Analysis (Classification)

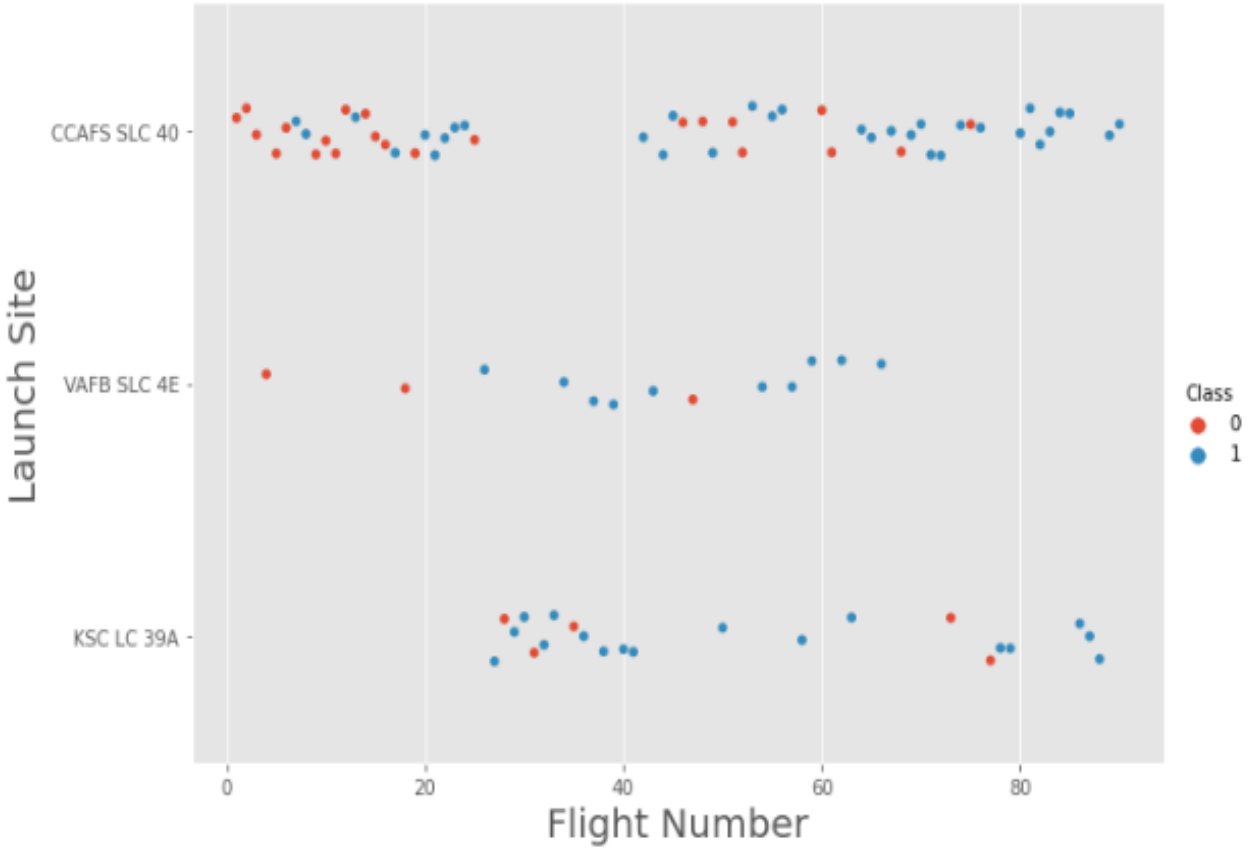
Github URL: SpaceX Machine Learning Prediction

1. Create a column for "Class".
2. Standardize the data.
3. Split into training and test set.
4. Find best Hyperparameter for SVM, Decision Trees, K-Nearest Neighbours and Logistic Regression.
5. Use test data to evaluate models based on their accuracy scores and confusion matrix.

Results

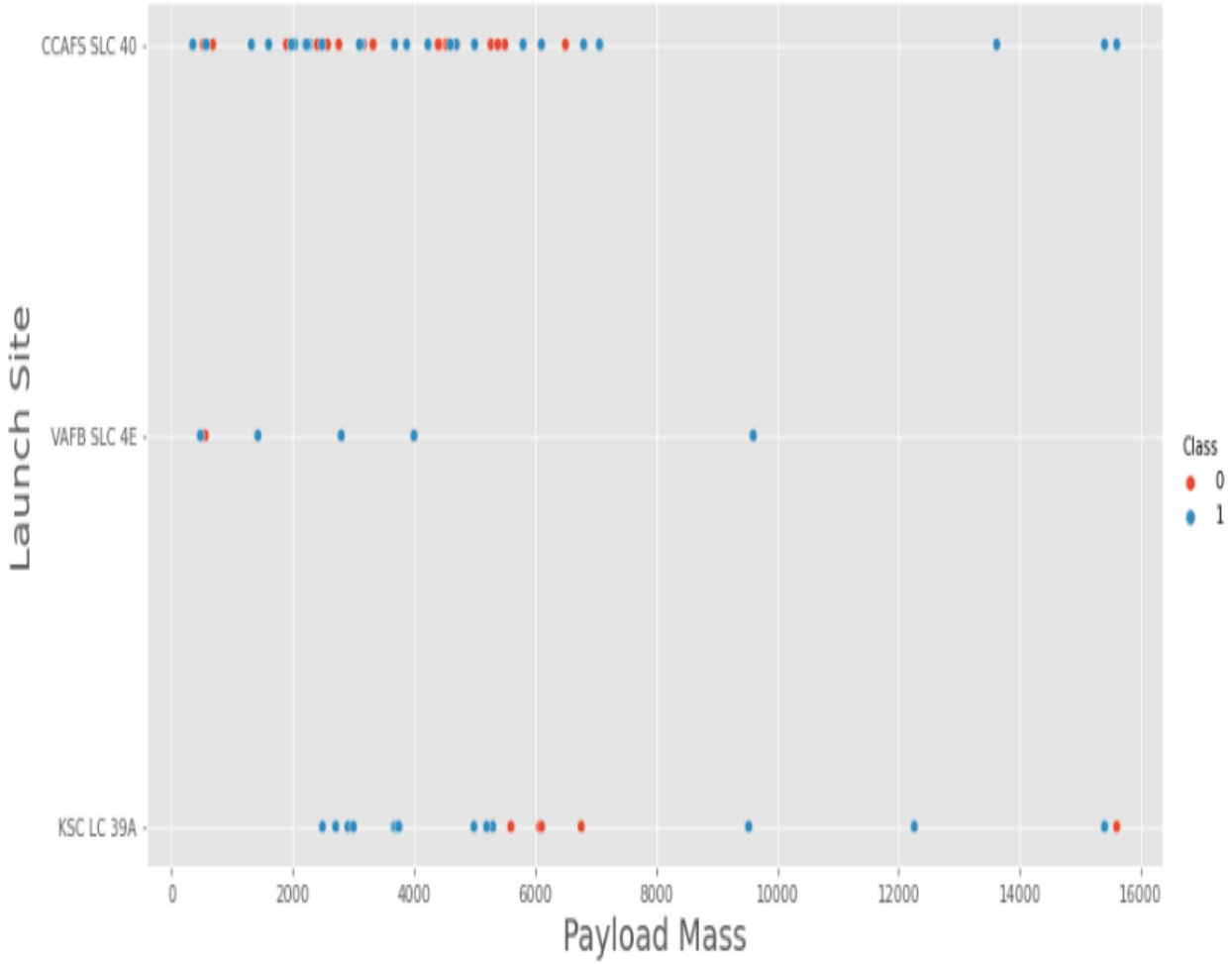
- The results of the exploratory data analysis revealed that the success rate of the Falcon 9 landings was 67%.
- The predictive analysis results showed that the Decision Tree algorithm was the best classification method with an accuracy of 94%

DISCUSSION - Flight Number vs. Launch Site



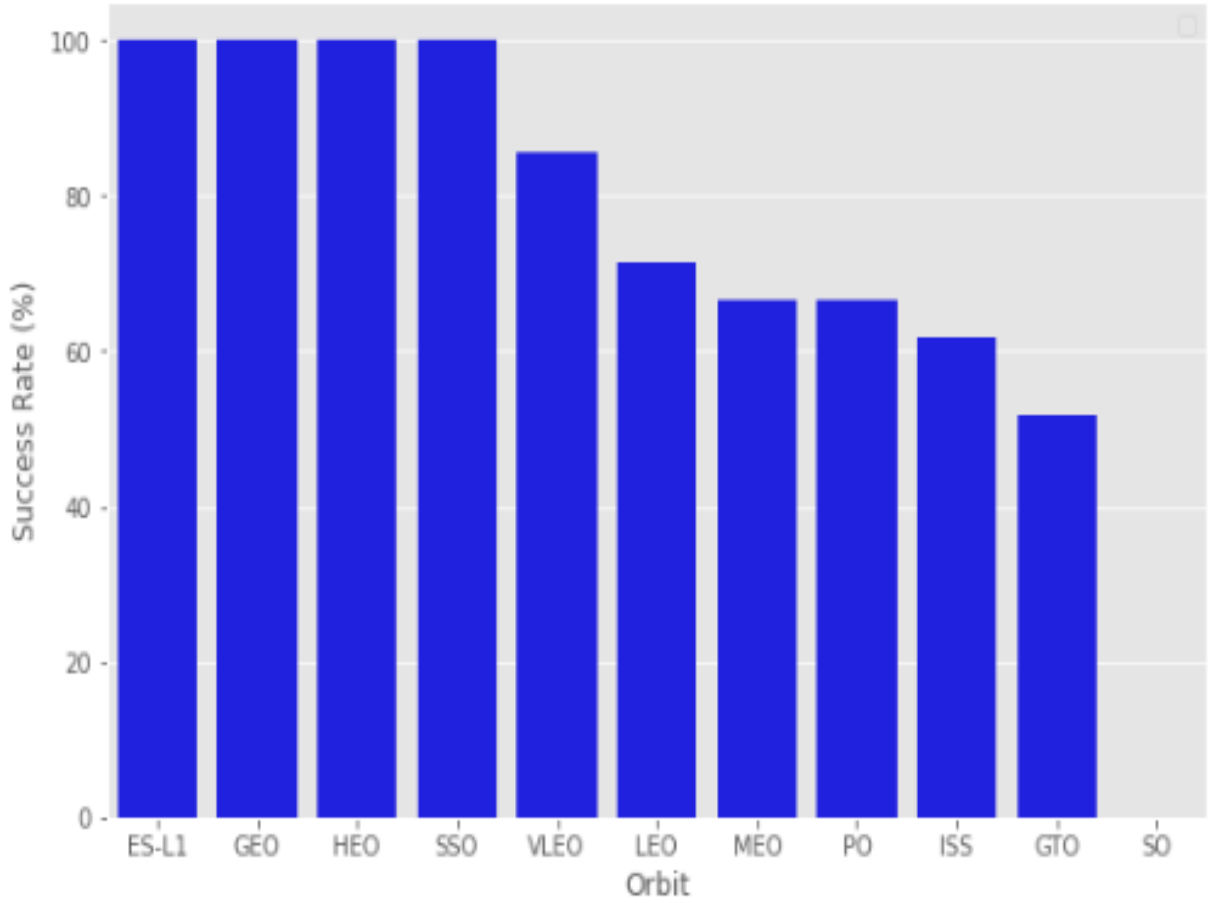
- The successful launches are shown by blue dots, whereas the failed launches are represented by red dots.
- This graph demonstrates that as the number of flights increased, so did the success rate.

Payload vs. Launch Site



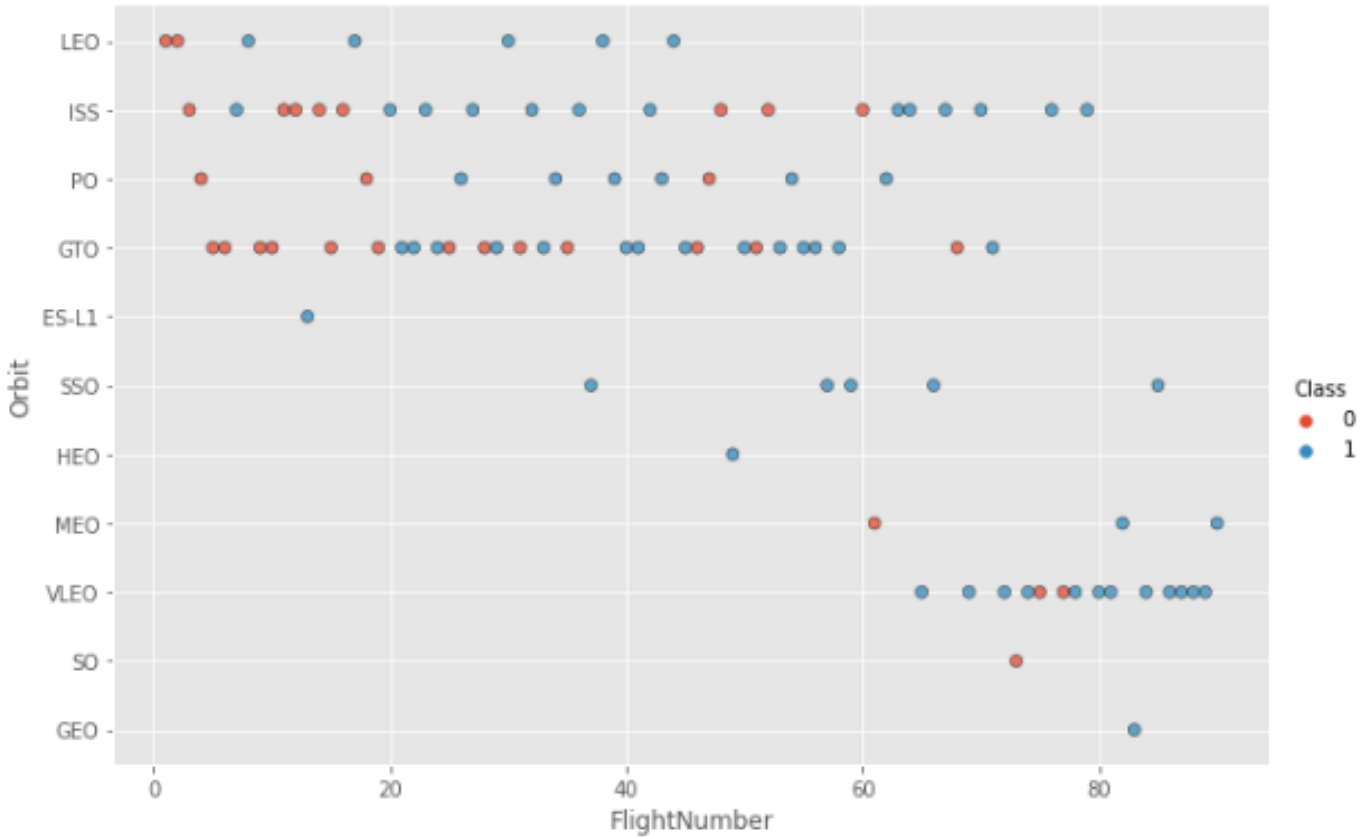
- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass
- No conclusions could be made using this metric because there appears to be a poor correlation between Payload and Launch Site.

Success Rate vs. Orbit Type



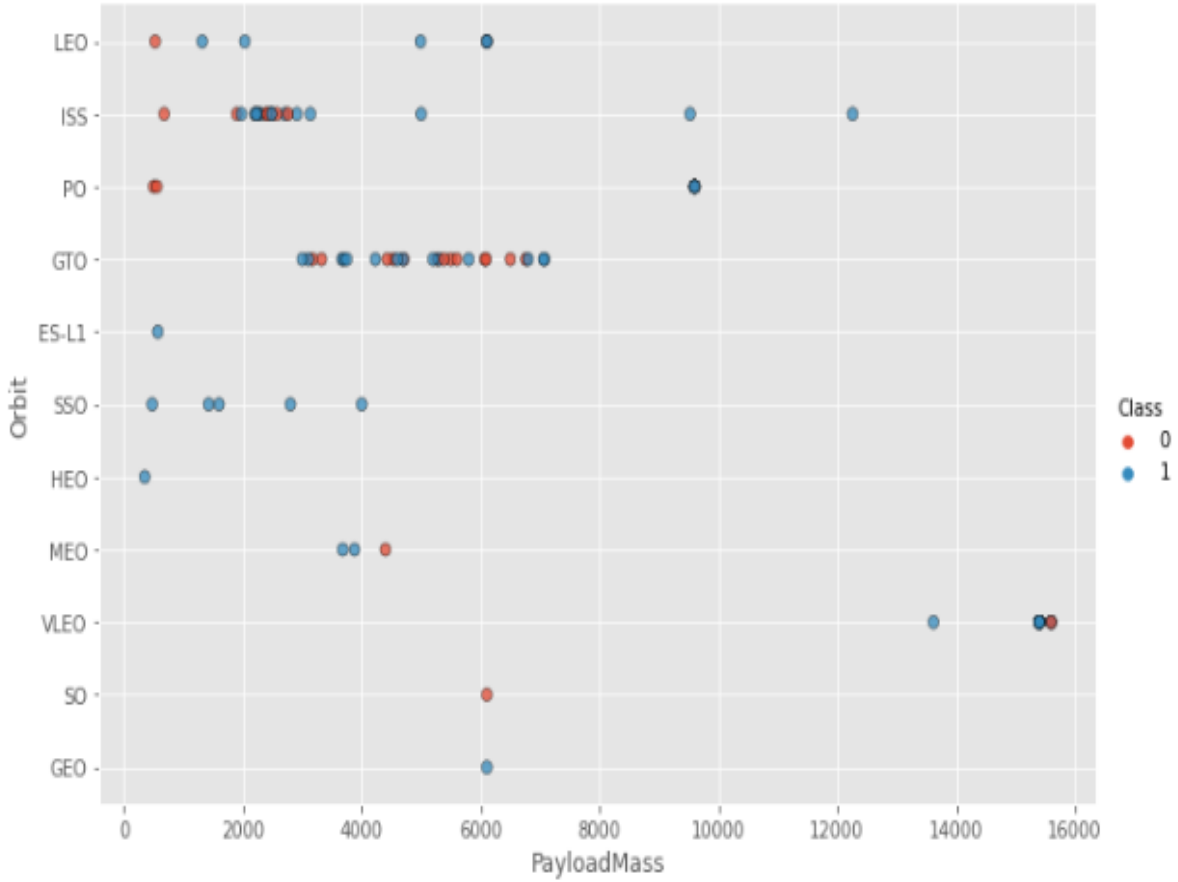
- Orbits SSO, HEO, GEO, and ES-L1 have 100% success rates.
- SO orbit did not have any successful launches.

Flight Number vs. Orbit Type



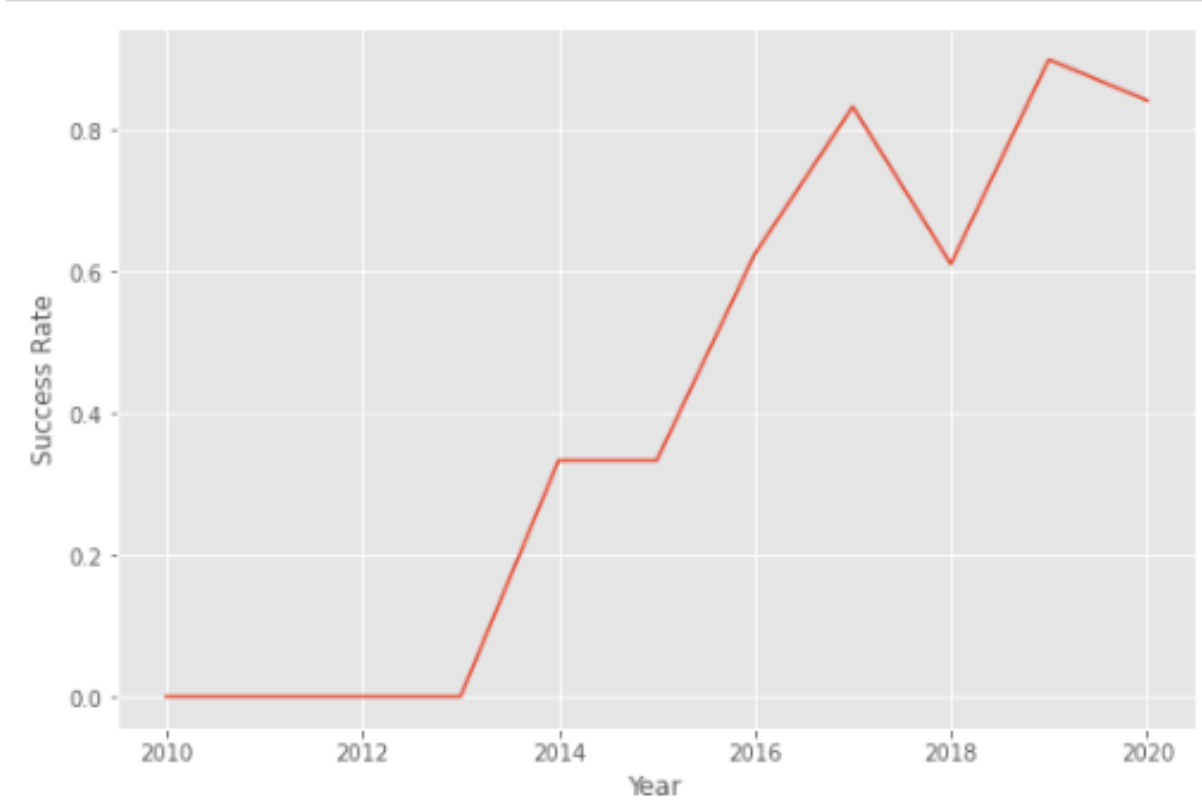
- In the LEO orbit, the success is positively correlated to the the number of flights.
- There seems to be no correlation between flight number in the GTO orbit.
- The SSO orbit has a 100% success rate however with fewer flights than the other orbits.
- Flights numbers greater than 40 have a higher success rate than flight numbers between 0-40.

Payload vs. Orbit Type



- The success rate rises in the PO, SSO, LEO, and ISS orbits as payload weight increases.
- Because both successful and unsuccessful launches are frequently observed, there does not appear to be a direct association between orbit type and payload mass for GTO orbit.

Launch Success Yearly Trend



- As the years go by, the chart's overall trend indicates an increase in landing success rate. However, both 2018 and 2020 see a decline.

Launch Sites

- The names of the launch sites are CCAFS LC-40, CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E .
- We used the DISTINCT clause was to return only the unique rows from the launch_site column.

Total Payload Mass

- The payload mass kg field was used to compute the overall payload that NASA boosters carried using the SUM() function.

total_payload_mass_kg
45596

Average Payload Mass by Falcon 9 v1.1

- The average payload mass carried by booster version F9 v1.1 was calculated using the AVG() tool.
- Results were filtered using the WHERE clause so that calculations were only carried out on booster versions if they were designated "F9 v1.1."

avg_payload_mass_kg
2928

First Successful Ground Landing Date

- The MIN(DATE) function was used to find the date of the first successful landing outcome on ground pad
- The WHERE clause ensured that the results were filtered to match only when the 'landing_outcome' column is 'Success (ground pad)'

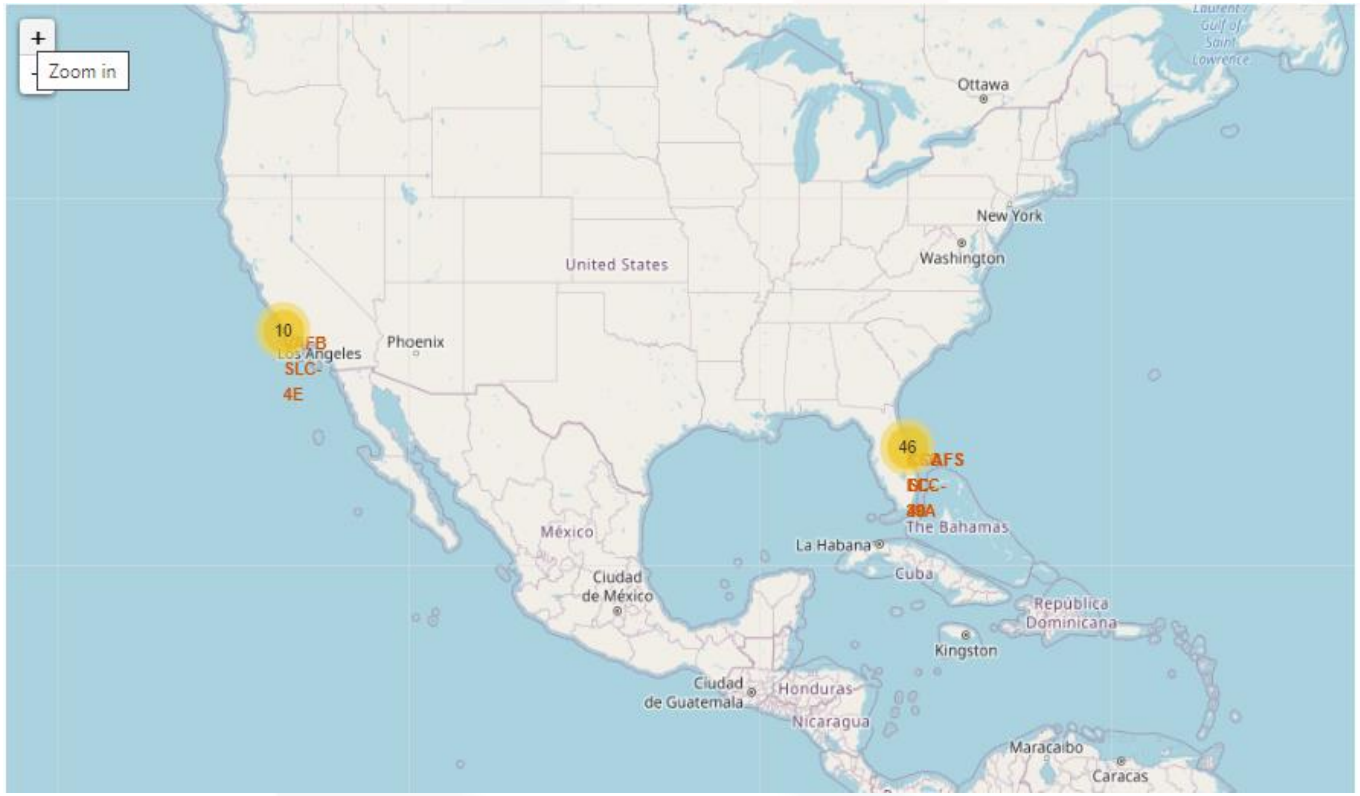
first_successful_landing_date
2015-12-22

Total Number of Successful and Failure Mission Outcomes

- With the help of the GROUPBY clause applied to the "mission_outcome" column, the COUNT() function is used to count the number of instances of various mission outcomes. The total number of mission outcomes—both successful and unsuccessful is returned.
- Out of 101 missions, 99 have resulted in successful missions.

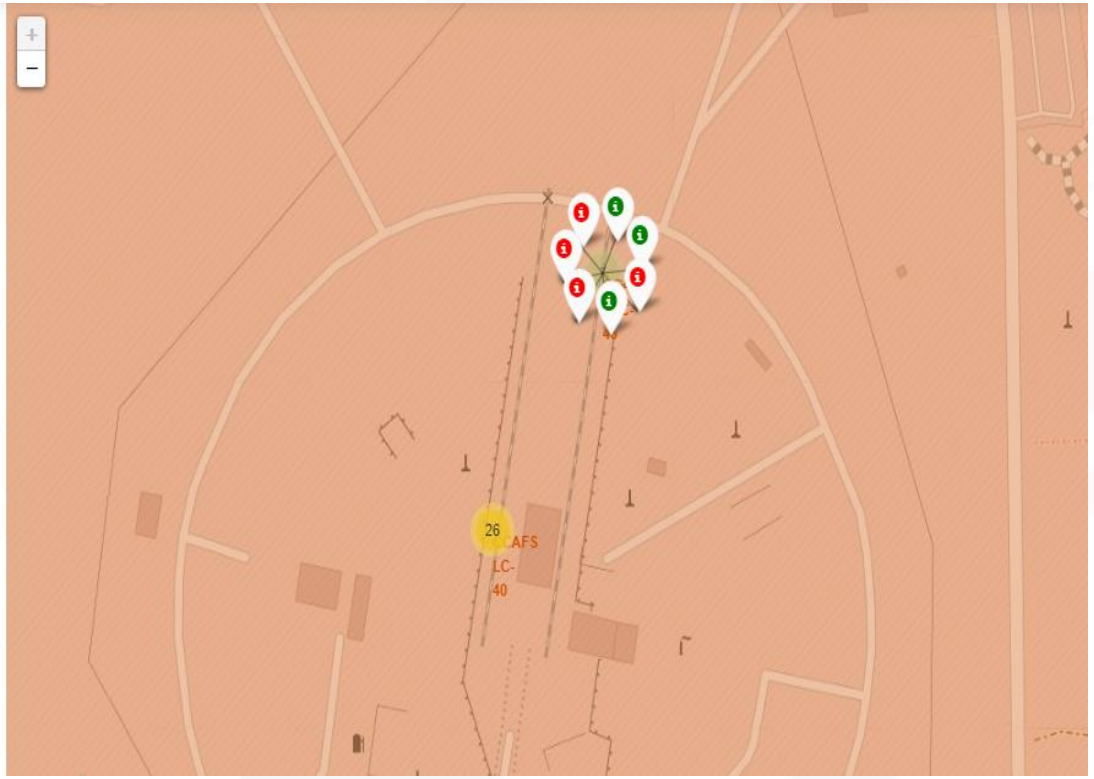
mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

SpaceX Launch Sites Locations



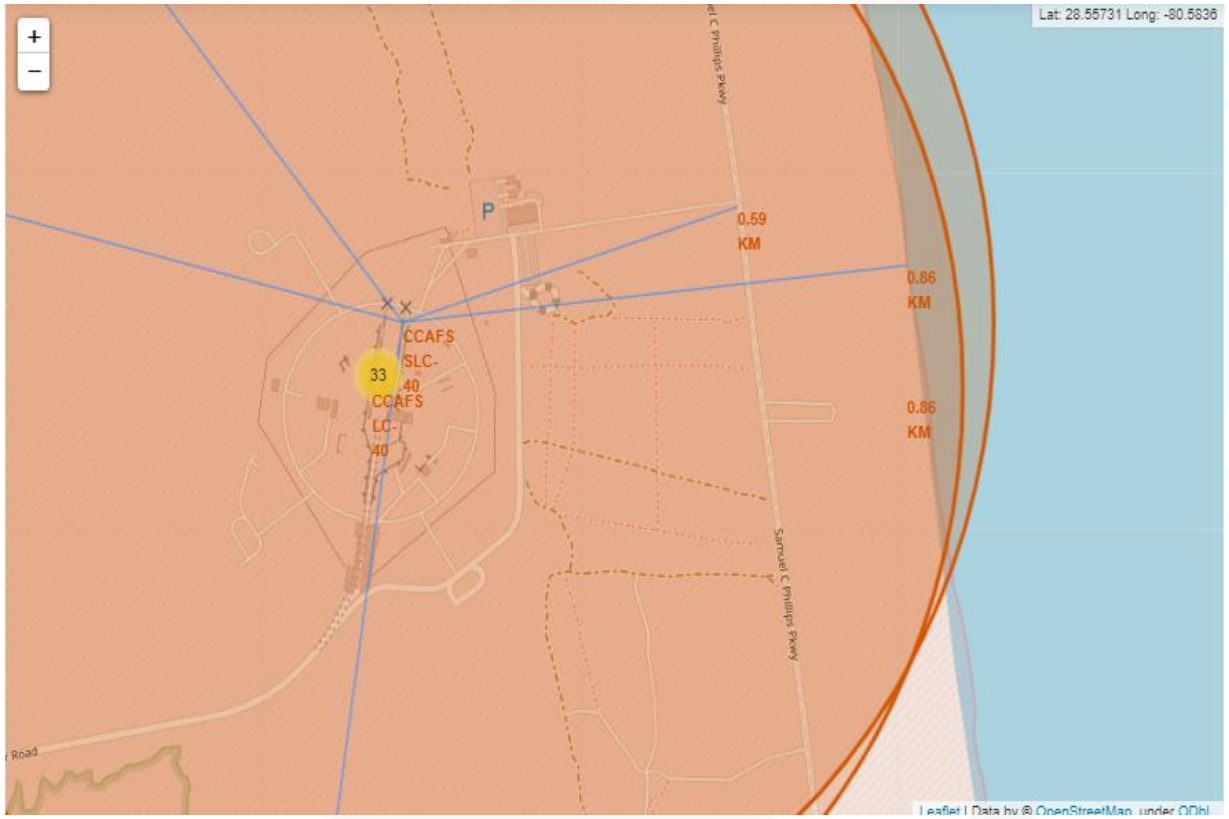
- The locations of all the SpaceX launch sites in the US are indicated by the yellow markers.

Success or Failure



- The launch site will show marker clusters of successful landings (green) or failed landings as we zoom in on a launch location (red).

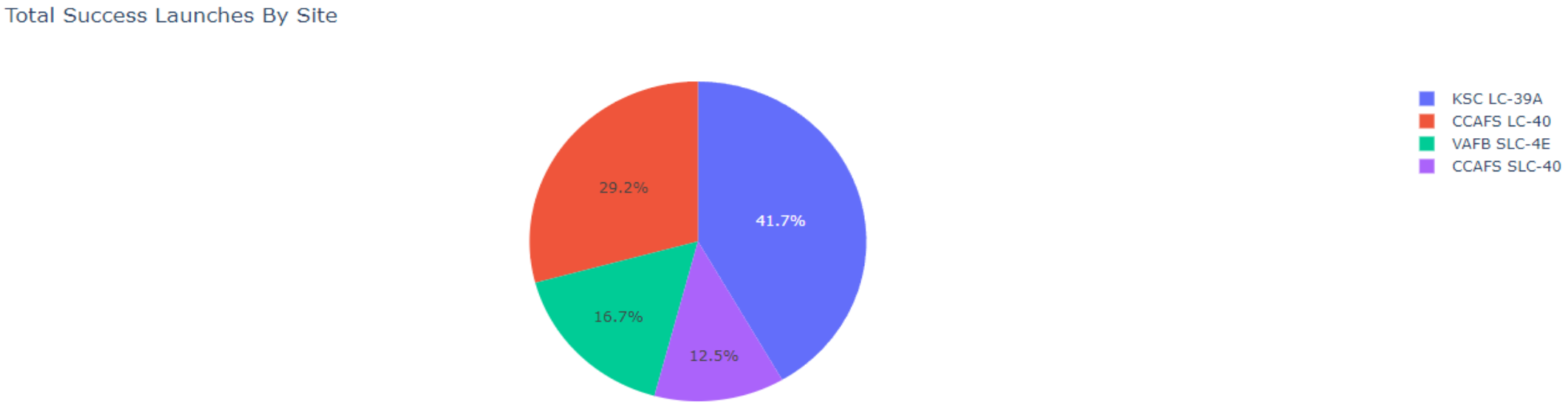
Launch Site Surroundings



- The generated map reveals that the chosen launch point is close to a highway for people and equipment transportation. For launch failure testing, the launch site is also close to the coast.

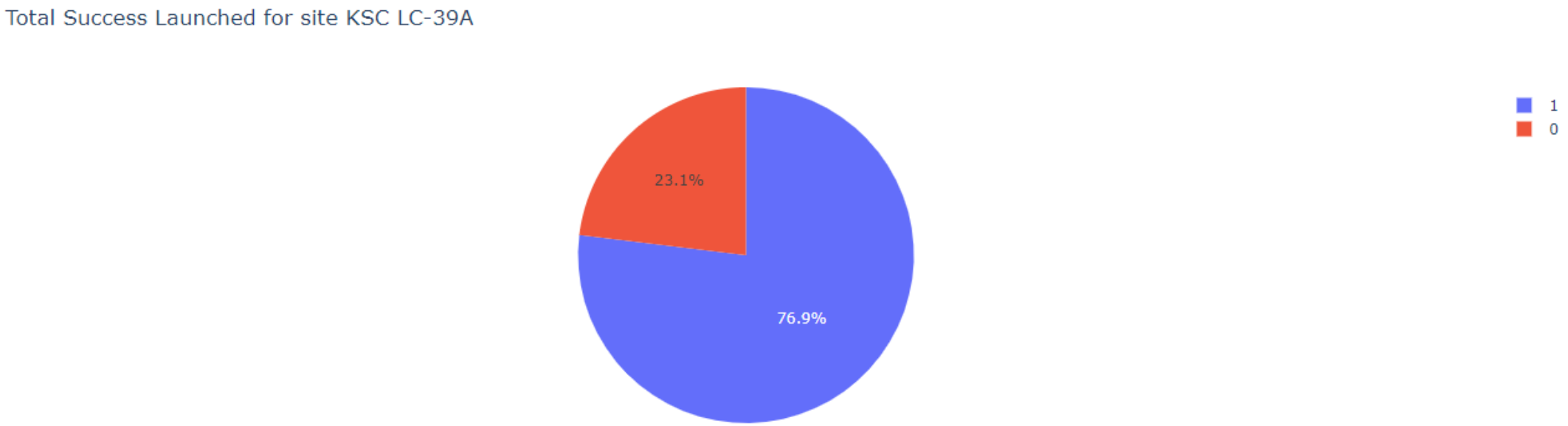
Total Successful Launches By Site

- The KSC LC-39A Launch site has the most successful launches with 10 in total.



Launch Site With Highest Success Ratio

- The KSLC-39A has the highest success rate with 76.9%.

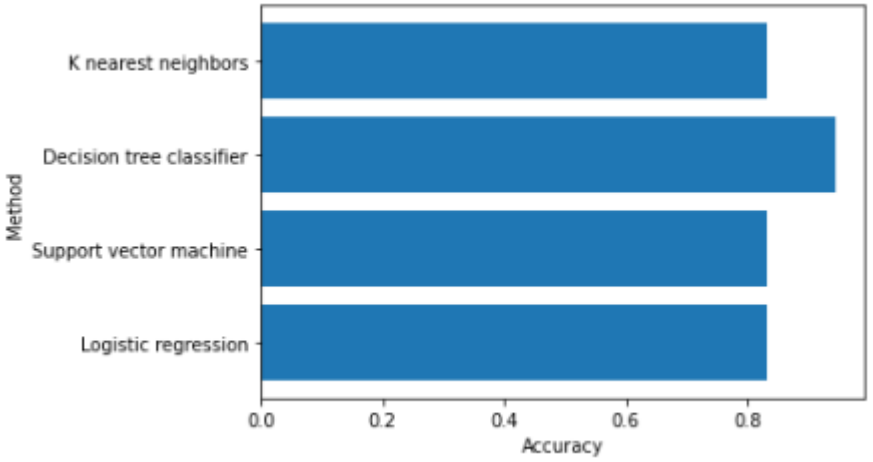


Payloads vs Launch Outcome

- Payloads between 0 and 2500 kg have a somewhat lower launch success percentage than payloads between 250 and 5000 kg. Actually, there isn't much of a distinction between the two.
- In both weight ranges, the v1.1 booster version had the highest effectiveness rate.

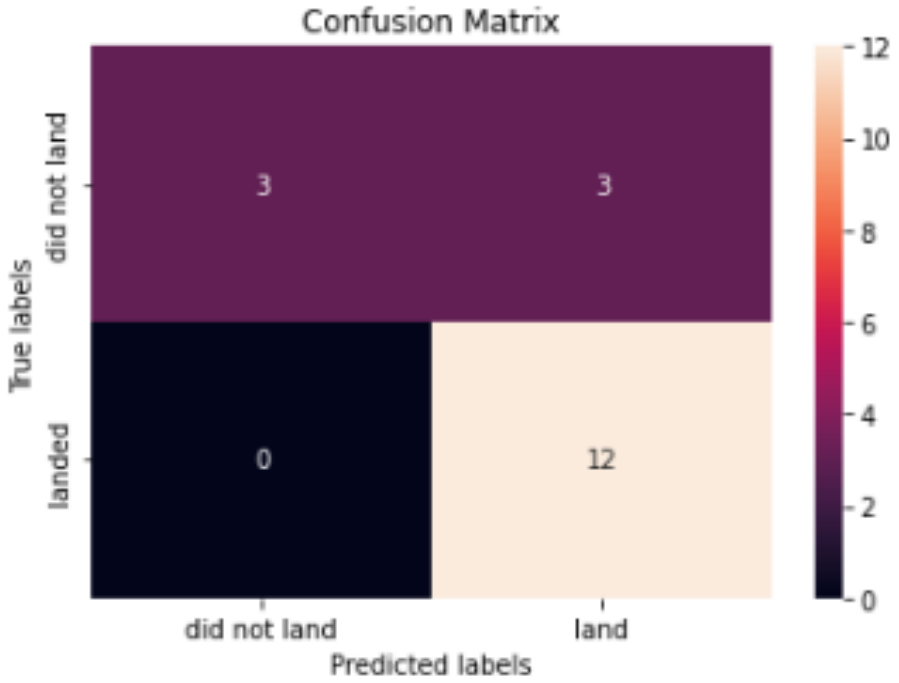
Classification Accuracy

The Decision Tree classifier had the best accuracy at 94%.



	method	accuracy
0	Logistic regression	0.833333
1	Support vector machine	0.833333
2	Decision tree classifier	0.944444
3	K nearest neighbors	0.833333

Confusion Matrix



- When the True label was success (True Positive), the model predicted 12 successful landings, and when it was failure, it predicted 3 unsuccessful landings (True Negative).
- The model also predicted 3 successful landings when the True label was unsuccessful landing (False Positive).

Conclusions

- The analysis revealed that as the success rate has increased over time, there is a positive correlation between the number of flights and success rate.
- The most successful launches occurred in orbits like SSO, HEO, GEO, and ES-L1.
- Payload mass can be related to success rate since lighter payloads have typically had better results than bigger payloads.
- The Decision Tree Classifier is the most accurate predictive model for this dataset, having a 94% accuracy rate.
- The launch sites are placed in a safe distance from cities but strategically close to roads and railroads for the transit of people and goods.

Appendix

- Github Repository: SpaceX Landing Prediction

Thank You!