IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Malak Almadhour>
<January 21, 2024.>

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Interactive Visual Analytics with Folium
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis result
  - Interactive analytics in screenshots
  - Predictive Analytics result

Introduction

Project background and context

SpaceX promotes Falcon 9 rocket launches on its website, offering a cost of 62 million dollars, significantly lower than the upwards of 165 million dollars charged by other providers. A key factor contributing to these cost savings is SpaceX's ability to reuse the first stage of the rocket. Consequently, determining the likelihood of a successful first stage landing is crucial in estimating the overall cost of a launch. This information becomes valuable in scenarios where alternative companies aim to compete with SpaceX for rocket launch contracts. The primary objective of this project is to develop a machine learning pipeline capable of predicting whether the first stage of the Falcon 9 rocket will successfully land.

Problems you want to find answers

What factors determine if the rocket will land successfully?
The interaction amongst various features that determine the success rate of a successful landing?
What operating conditions needs to be in place to ensure a successful landing program?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was collected using SpaceX API and web scraping (beautiful soap) from Wikipedia.
- Perform data wrangling
  - The data collected in form of Jason object and HTML tables and then we converted the data into pandas data frame.
- 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

- The data was gathered from the Space X REST API and web scraping from wiki pages.

Space X REST API endpoint	Get requests using the Requests library	Get past Launch data as JSON object	Convert the JSON to a df
Web scribing Flacon 9 records	Use beautiful soap to web scape HTML tables	Parse data from tables	Convert tables a dataframe

Data Collection – SpaceX API

Collect and make sure the dat in the correct format from an API

Space X REST API endpoint

Get requests using the Requests library

Extract information about booster name launch site, payload mass and landing site

Convert the past launch data as JSON object

Data wrangling

Deal with missing values

Filter the data frame to only include falcon 9

Convert the JSON to a df

https://github.com/Malakalmadhor/Final-project.-/blob/514329100e694bd847760ff6d1348f7490a835c1/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

request the falcon 9
launch wiki pages
From its URL

beautiful Soup object
from the response

Extract all column/
variable names from the
HTML table headers

Create a data frame by
parsing the launch HTML
tables

Data wrangling

https://github.com/Malakalmadhor/Final-project.-/blob/514329100e694bd847760ff6d1348f7490a835c1/jupyter-labs-webscraping.ipynb

Data Wrangling

Perform explanatory data analysis EDA to find patterns in the data and determine what would be the label for train supervised models

identify the missing values

identify which column are numerical and categorical

Calculate the number of launches on each site

Create a landing outcome label from outcome column

Calculate the number and occurrence of mission outcome per orbit type

calculate the number and occurrence of each orbit

https://github.com/Malakalmadhor/Final-project.-/blob/514329100e694bd847760ff6d1348f7490a835c1/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Charts that were plotted
Catplot to visualize the relationship between flight number and play load
Catplot lot visualizer relationship between flight number and launch site
Catplot to visualize the relationship between payload and launch site
Bar chart to visualize the relationship between success rate each orbit type
Catplot to visualize the relationship between flight number and orbit, type
catplot visualize a relationship between payload and orbit type
line chart to visualize the launch success yearly trend

https://github.com/Malakalmadhor/Final-project.-/blob/df01828c003f05d37be91e39cbec855a28b3b0cf/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

SQL queries performed

- Display the names of the unique launch sites in the space mission:
  SELECT DISTINCT (launch_site) FROMSPACEXTBL;

- Display 5 records where launch sites begin with the string 'CCA:
  SELECT * FROM SPACEXTBL WHERE launch_site LIKE 'CCA%' LIMIT 5;

- Display the total payload mass carried by boosters launched by NASA (CRS):
  SELECT SUM(payload_mass_kg_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL WHERE customer='NASA (CRS);

- Display average payload mass carried by booster version F9 v1.1:
  SELECT AVG(payload_mass_kg_) AS AVG_PAYLOAD_MASS FROM SPACEXTBL WHERE booster_version=*F9 v1.1';

- List the date when the first successful landing outcome in ground pad was achieved:
  SELECT MIN(DATE) AS first_successful_landing FROM SPACEXTBL WHERE (landing_outcome) = 'Success(ground pad)';

https://github.com/Malakalmadhor/Final-project.-/blob/58b42813e3be09dbc20a81fa3dd4ad44fa8dcf74/jupyter-labs-eda-sql-coursera_sqllite.ipynb

EDA with SQL

SQL queries performed

SQL queries performed:
• List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000:
SELECT booster_version, payload_mass_kg_, landing_outcome FROM SPACEXTBL WHERE landing_outcome=Success (drone ship)' AND (payload_mass_kg_ BETWEEN 4000 AND 6000);

• List the total number of successful and failure mission outcomes:
SELECT mission_outcome, COUNT(mission_outcome) AS TOTAL FROM SPACEXTBL GROUP BY mission_outcome;

• List the names of the booster_versions which have carried the maximum payload mass. Use a subquery:
SELECT DISTINCT(booster_version), (SELECT MAX(payload_mass_
kg_) AS "maximum_payload
_mass"FROM SPACEXTBL) FROM SPACEXTBL LIMIT 5

https://github.com/Malakalmadhor/Final-project.-/blob/58b42813e3be09dbc20a81fa3dd4ad44fa8dcf74/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

Summary of map objects that were created and added to the Folium map

folium.Circle and folium.Marker to add a highlighted circle area with a text label on a specific coordinate for each launch site on the site map.

MarkerCluster object for simplify a map containing many markers having the same coordinate.

MousePosition on the map to get coordinate for a mouse over a point on the map.

folium.PolyLine object to draw a line between a launch site to its closest city, railway and highway.

- https://github.com/Malakalmadhor/Final-project.-/blob/78507b3ef5574de54fc50f132bfaeef3562e0f1f/lab_jupyter_launch_site_location.jupyterlite.ipynb

Build a Dashboard with Plotly Dash

Summary of plots/graphs and interactions that were added to the dashboard to perform interactive visual analytics on SpaceX launch data in real-time.

This dashboard application contains input components such as a dropdown list and a range slider to interact with a pie chart and a scatter point chart.

A launch Site Drop-down Input Component.
There are four different launch sites and a dropdown menu let us select different launch sites.

A callback function to render success-pie-chart based on selected site dropdown.
The general idea of this callback function is to get the selected launch site from site-dropdown and render a pie chart visualizing launch success counts.

A range Slider to Select Payload.
The Slider is to be able to easily select different payload range and see if we can identify some visual patterns.

https://github.com/Malakalmadhor/Final-project.-/blob/f380f24a3f45e010863ea6847746e232bb6dda34/Build%20a%20Dashboard%20with%20Plotly%20Dash

Predictive Analysis (Classification)

Summary of the model development process used to predict if the first stage will land given the data from the preceding labs.

Creation of a NumPy array from the column Class in data.
Data standardization.

Use of the function train_test_split to split the data X and Y into training and test data.

Searching for the best Hyperparameters for Logistic Regression, SVM, Decision Tree and KNN classifiers.

Searching for the method that performs best using test data

- https://github.com/Malakalmadhor/Final-project.-/blob/07733b248329082fe8f2c0e8de75b85d8197812d/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Predictive Analysis (Classification)

Load of the data "dataset_part_2.csv" and "dataset_part_3.csv"

Use of "dataset_part_2.csv" for creation of variable Y from the column Class

Use of "dataset_part_3. CSV" the features_one_hot dataframe for creation of variable X

train_test_split to split the data into training and test data

Creation of a Logistic Regression, SVM, Decision Tree, KNN and GridSearchCV objects

Best parameters, accuracy and confusion matrix

Selection of method That performs best

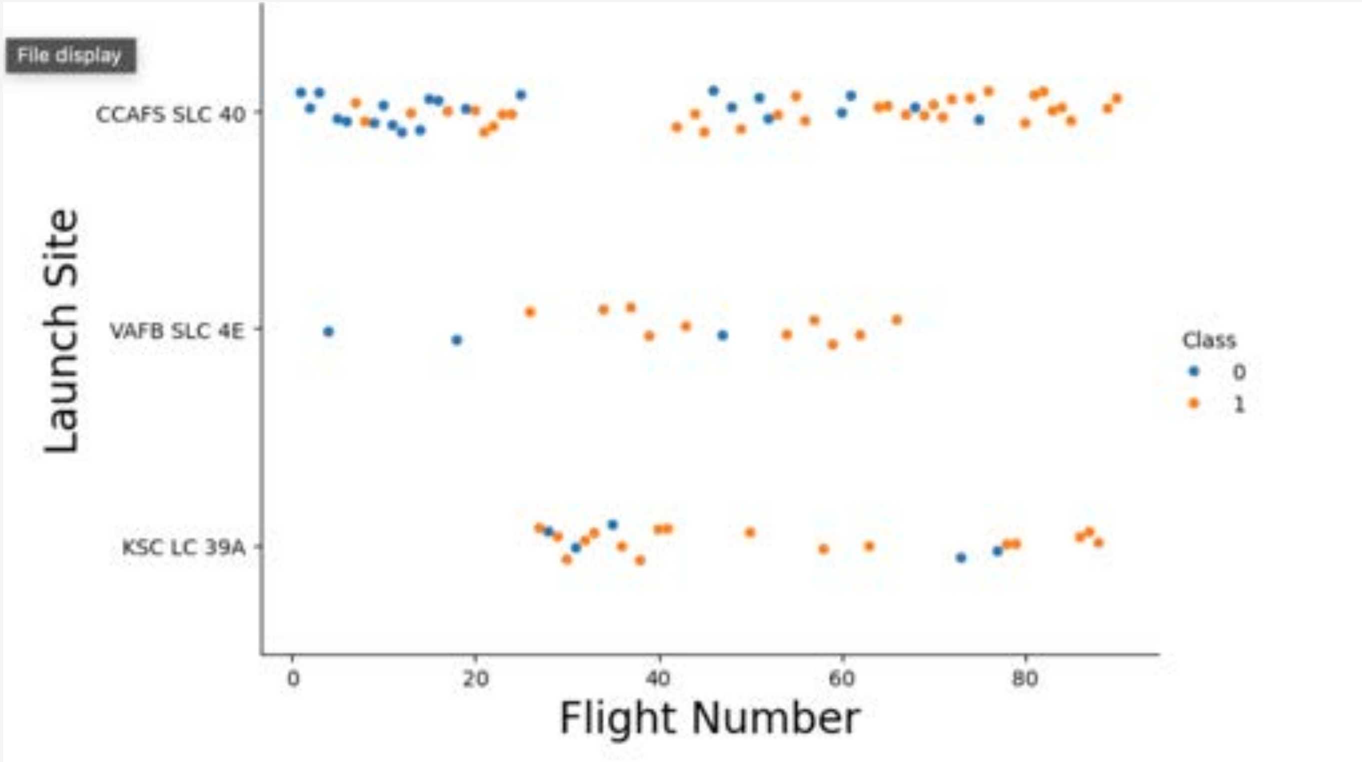
Results

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

Section 2

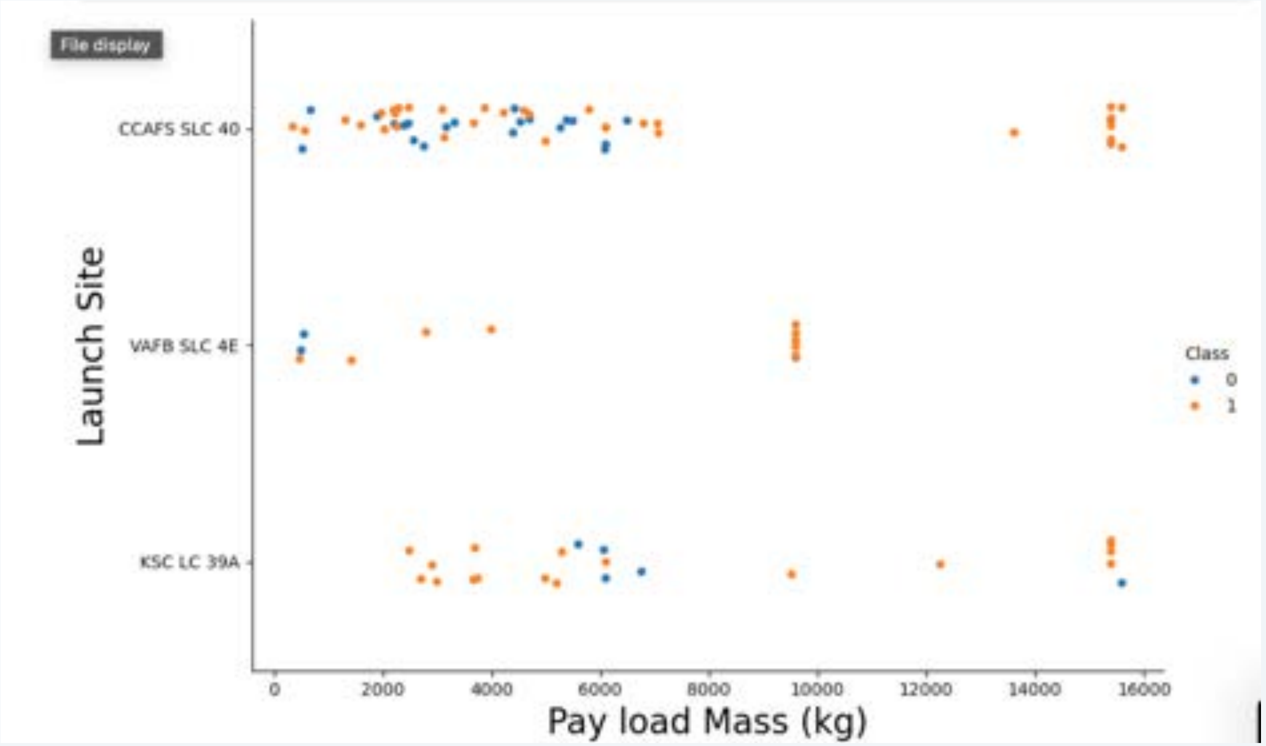
Insights drawn from EDA

Flight Number vs. Launch Site



With time the successful rate has increased for every Launch Site, especially for CCAFS SLC 40, where are concentrated the majority of the launches.
VAFB SLC 4E and KSC LC 39A has a higher successful rate but represents one third of the total launches.

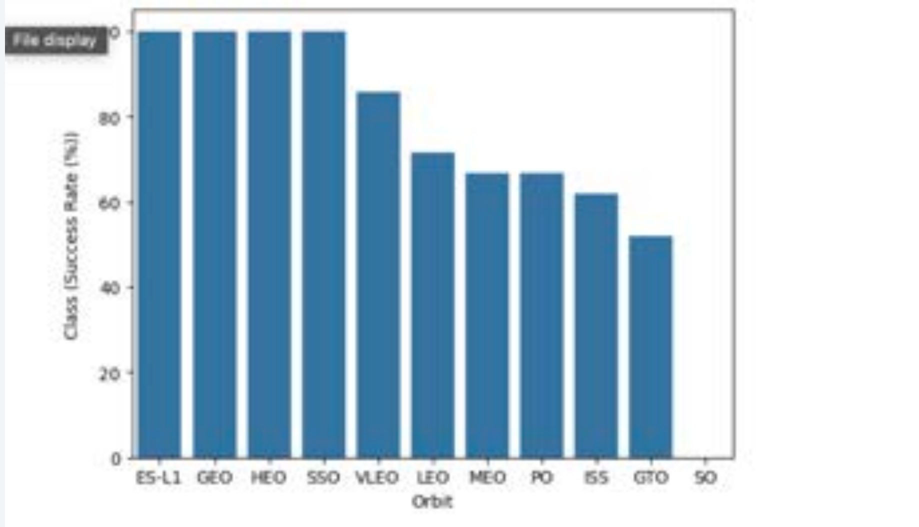
Payload vs. Launch Site



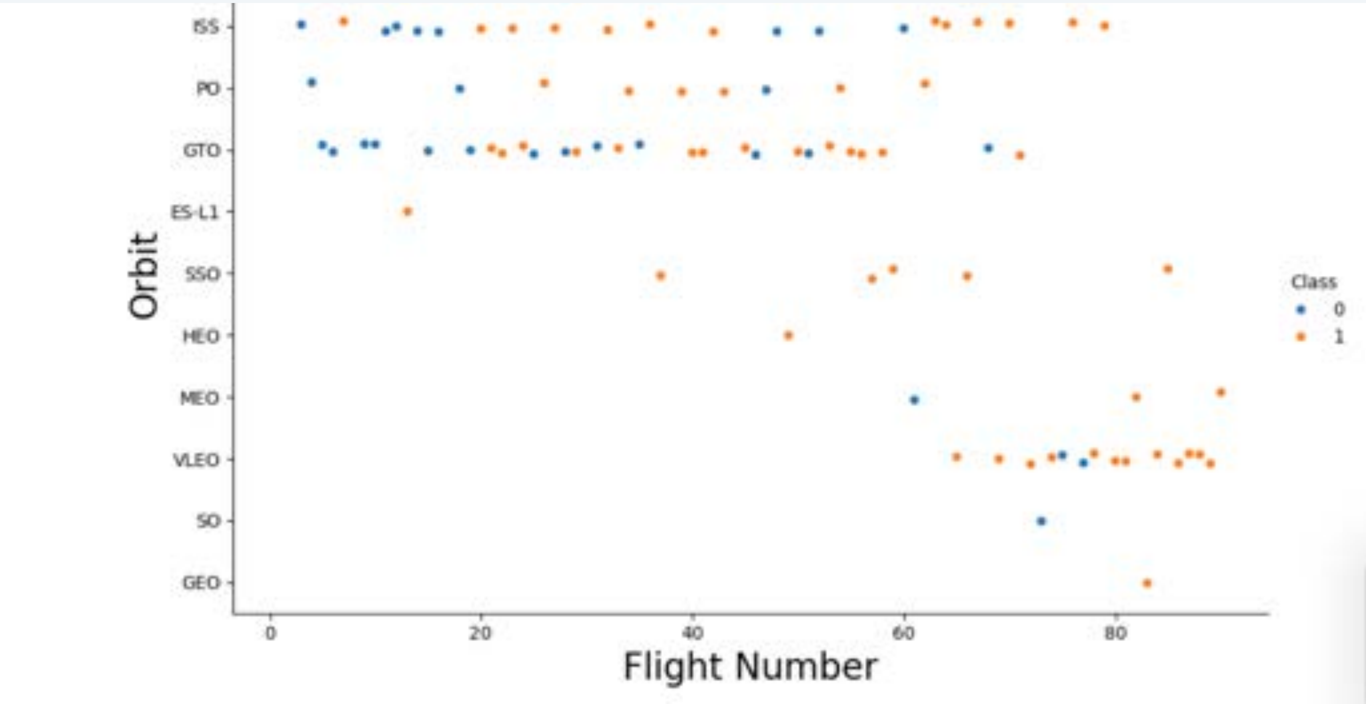
In VAFB-SLC launch site there are no rockets launched for heavy payloadmass (greater than 10000 kg).
In KSC LC launch site there are no rockets launched for lower payloadmass (less than 2500kg).
CCAFS SLC has launched rockets less than 7500kg and more than 13000kg payloadmass but not in between.

Success Rate vs. Orbit Type

The first 4 Orbit types has the best successful rate.
But how many attempts are per orbit type?
The bar chart must be interpreted with the number of launches per orbit type.

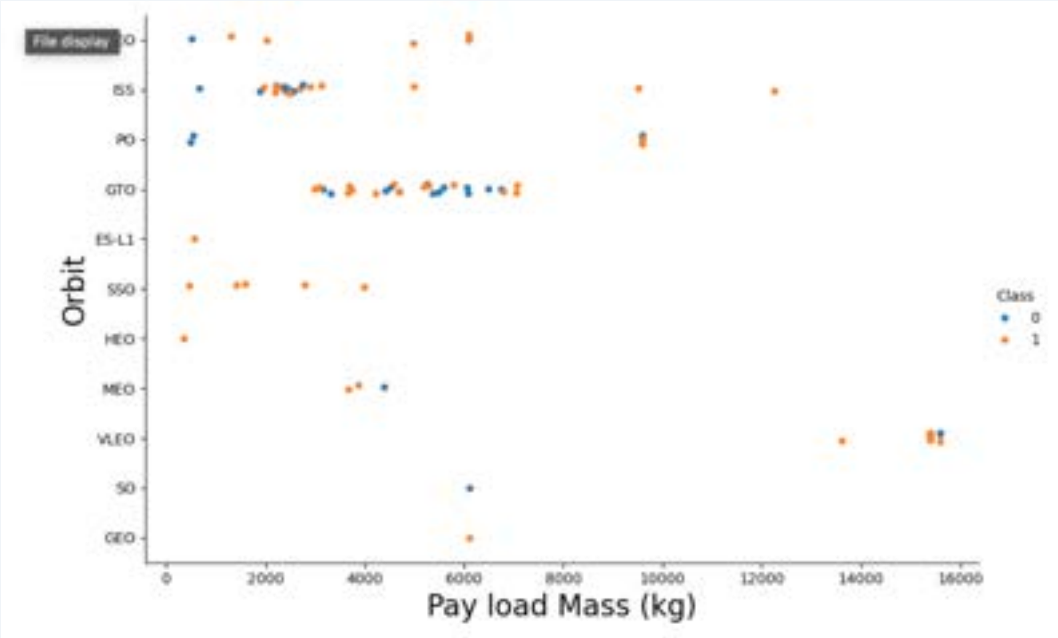


Flight Number vs. Orbit Type



As expected, there are more failures at the beginning of the series of launches, but, after the first 40 launches, the ratio improves by reducing the 50 percent of unsuccessful landings.
GTO and ISS orbits has the higher concentration of launches with the lowest ratio of successful landings.
The orbits with higher successful rate, has one or just a few number of launches.

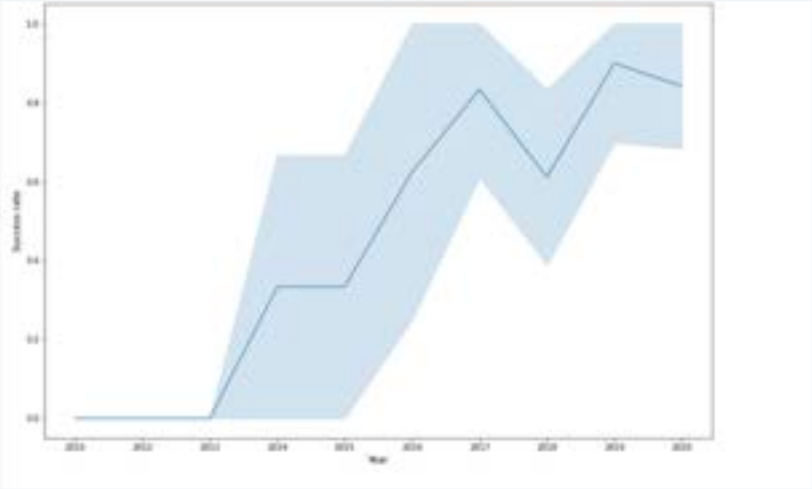
Payload vs. Orbit Type



Exists a visible limit of Payload around 7600 kg. Less than 10 launches exceed that limit.
With heavy payloads the successful landing rate are more for Polar, LEO and ISS.
However for GTO, we cannot distinguish this well as both, positive landing rate and negative landing are both there here.

Launch Success Yearly Trend

The success rate since 2013 kept increasing until 2020.



All Launch Site Names

The four unique launch sites in the space mission.
I have used "DISTINCT" statement to find the unique values in the launch site column.

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

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

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with the string 'CCA'. The query uses WHERE, LIKE and LIMIT.

Done.

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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 total payload mass carried by boosters launched by NASA (CRS) using SUM function and WHERE clause.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) as "Total Paylo
```

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

Total Payload Mass(Kgs)	Customer
45596	NASA (CRS)

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 using AVG function.

avg_payload_mass
2928

First Successful Ground Landing Date

The date when the first successful landing outcome in ground pad was achieved using MIN function.

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000, combining WHERE clause with AND operator.

booster_version	payload_mass__kg_	landing_outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes. The query uses a combination of COUNT function with GROUP BY statement.

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

Boosters Carried Maximum Payload

The names of the booster_versions which have carried the maximum payload mass. Using a subquery.

booster_version	maximum_payload_mass
F9 B4 B1039.2	15600
F9 B4 B1040.2	15600
F9 B4 B1041.2	15600
F9 B4 B1043.2	15600
F9 B4 B1039.1	15600

2015 Launch Records

The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

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. The query uses COUNT, WHERE, BETWEEN and GROUP BY.

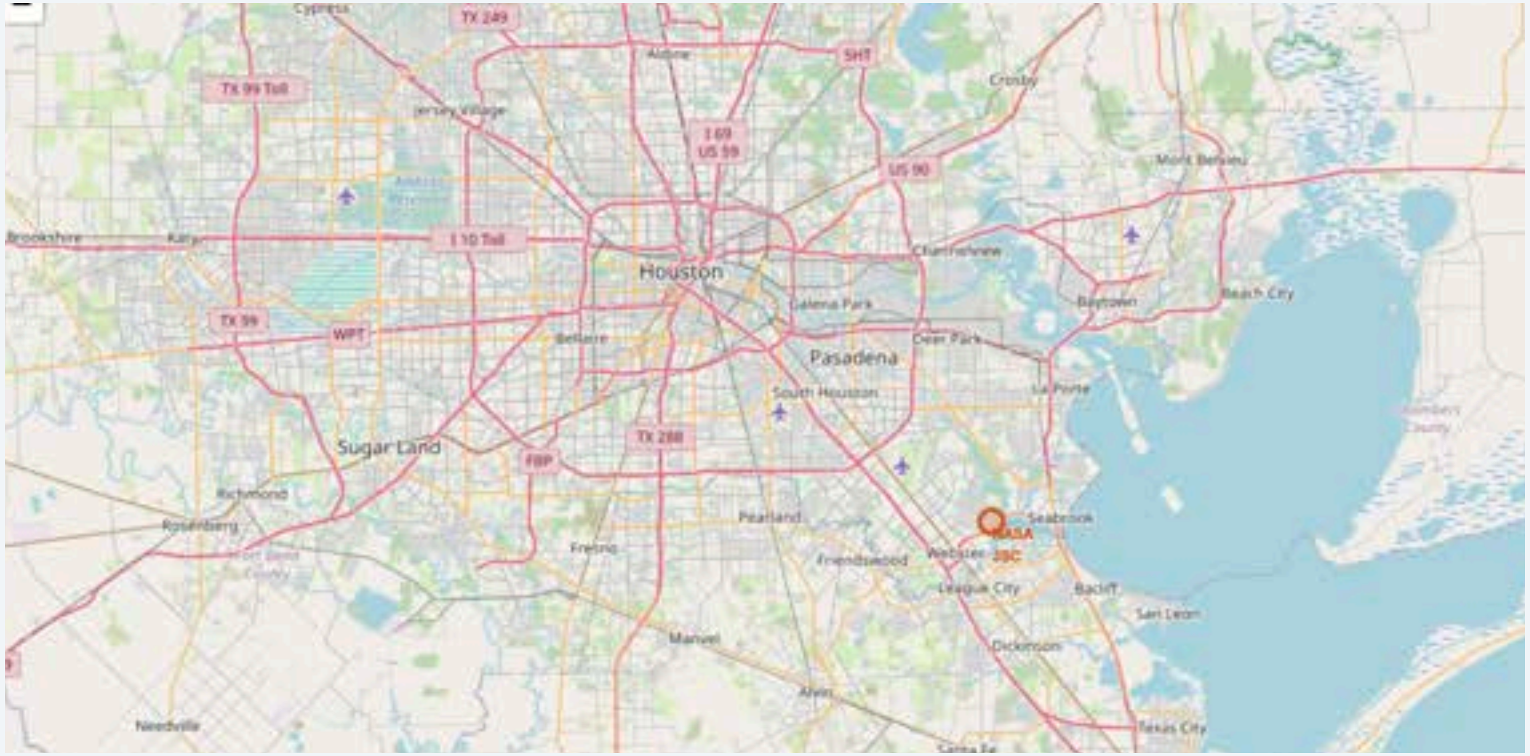
landing_outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

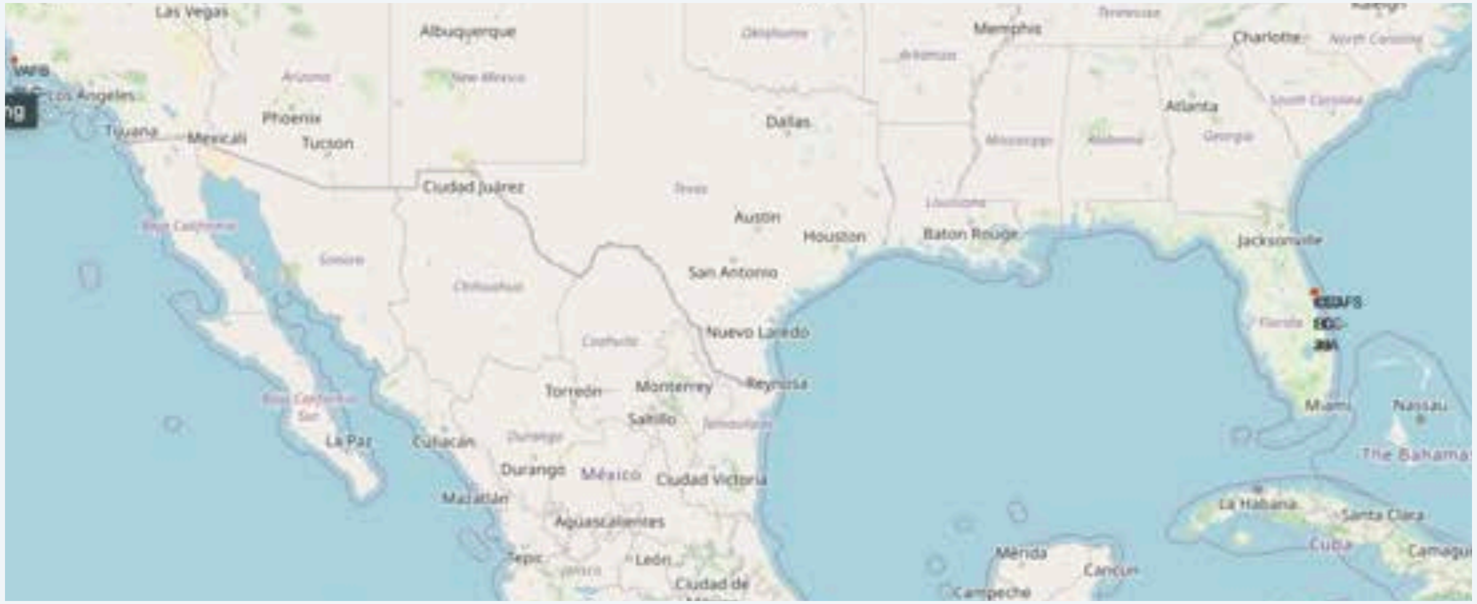
All launch sites

All launch sites are in very close proximity to the coast and into restricted areas.



Success/Failed Launches For Each

The first map shows clusters for every launch site, the second shows a green marker if a launch was successful, and a red marker if a launch was failed.



A Launch Site And Its Proximities

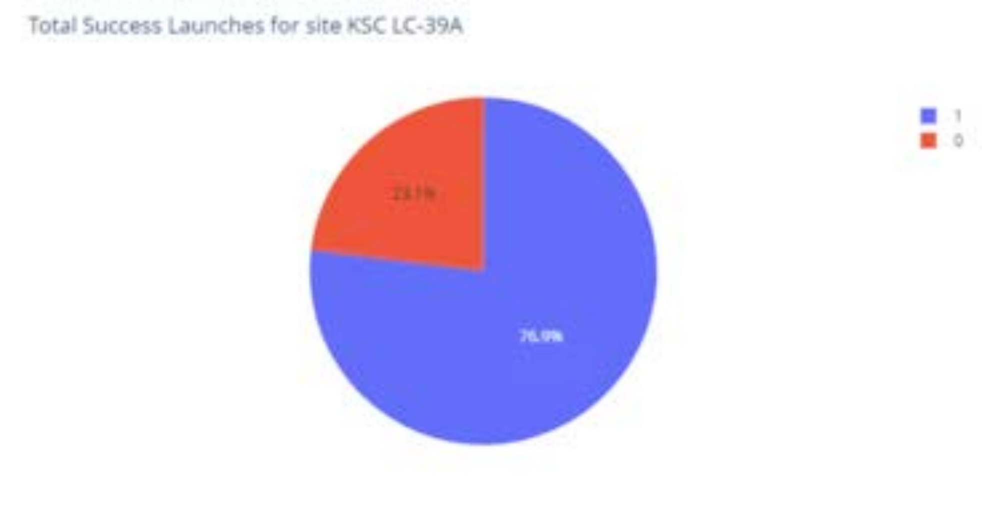
Launch sites are near to railways, roads, highways and coastline.
I understand that it is not just for easy supply or access but, for maintain a safe distance with near cities

Section 4

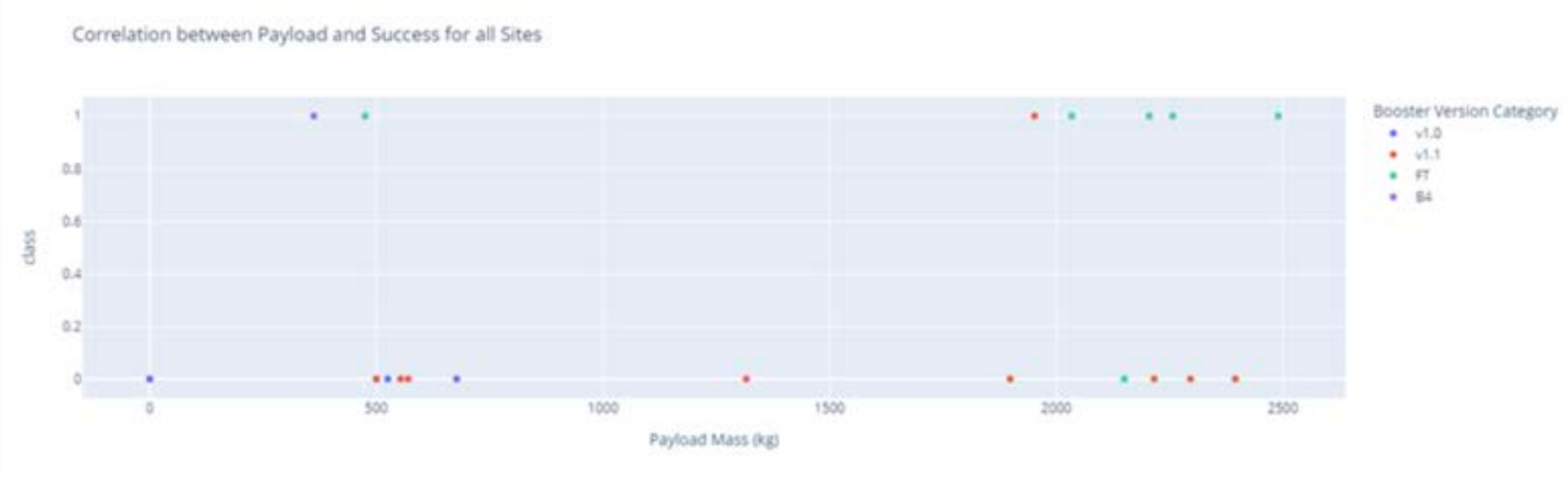
Build a Dashboard with Plotly Dash

Total Success Launches By Site

KSC LC-39A



Payload vs. Launch Outcome

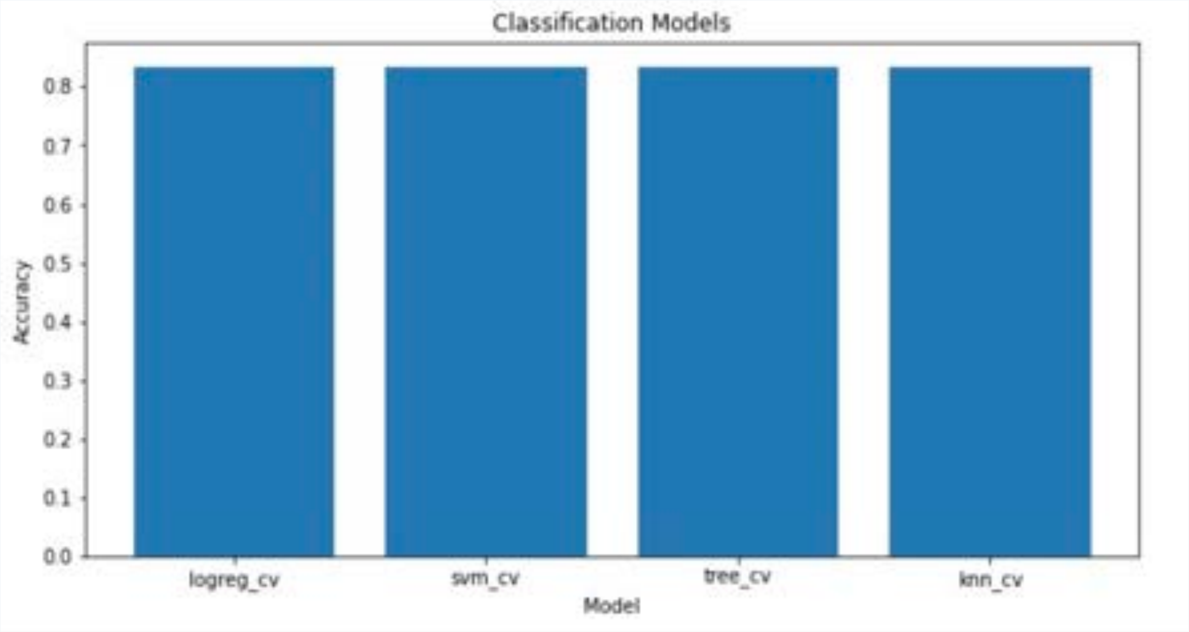


Section 5

Predictive Analysis (Classification)

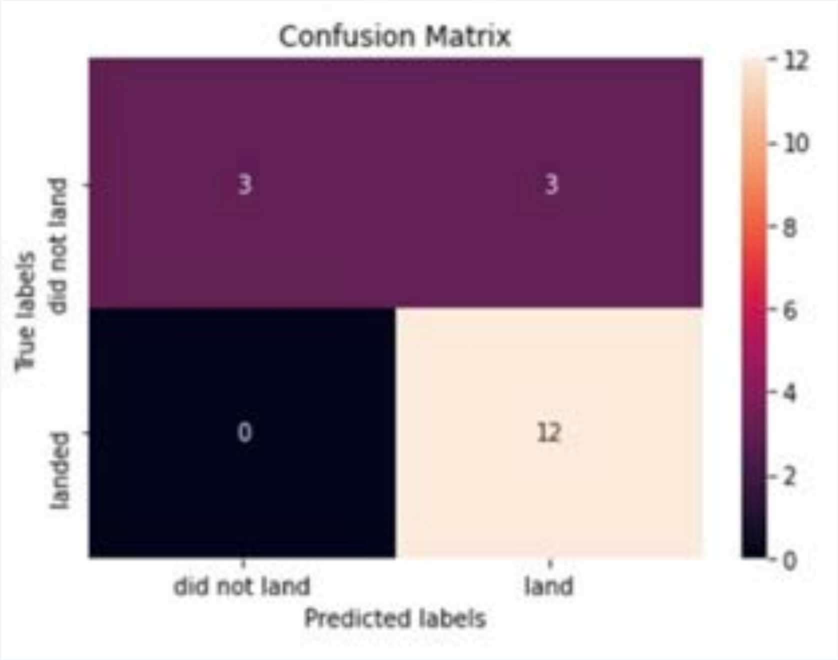
Classification Accuracy

The accuracy is the same for all models.



Confusion Matrix

The confusion matrix is the same for all models.



Conclusions

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.
- As all the algorithms are giving the same accuracy, they all perform practically the same.
- By using our machine learning model, we can predict if the first stage of our competitor will land and determine the cost of a launch.

Appendix

For notebooks, datasets and scripts, follow this GitHub repository link:

https://github.com/Malakalmadhor/Final-project.-.git

Thank you!