IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Joseph Emmanuel
14 Jul 2024

Outline

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

Executive Summary

Objective: Predict the success of Falcon 9 first stage landings to reduce launch costs.

Methodology:

- **Data Source:** SpaceX API
- **Tools Used:**
  - Matplotlib and Dash Plotly for visualizations
  - Folium for geographical mapping
  - SQL for targeted data analysis

Key Findings:

Important factors for successful landings:

- **Launch Site Location**
- **Payload Mass**
- **Orbit Type**

Predictive Models and Results:

- **Decision Trees:** Highest accuracy at 94.4%
- **Logistic Regression, SVM, KNN:** Each achieved 83.3% accuracy

Conclusion: Decision Trees provide the most accurate predictions for Falcon 9 first stage landings, supporting more cost-effective and reliable rocket launches for SpaceX.

Introduction

Project Background and Context

The commercial space age has revolutionized space travel, making it more accessible and affordable. Leading companies like Virgin Galactic, Rocket Lab, Blue Origin, and particularly SpaceX, have driven this transformation.

Key Achievements of SpaceX:

- **ISS Missions:** Regularly sends spacecraft to the International Space Station.
- **Starlink:** Developed a satellite constellation providing global internet coverage.
- **Manned Missions:** Successfully conducts manned space missions.

A major factor behind SpaceX's success is the reuse of the Falcon 9 rocket's first stage, which drastically lowers launch costs. Competitors charge upwards of $165 million per launch, while SpaceX offers the Falcon 9 for approximately $62 million, thanks to its reusable first stage.

Introduction

Objectives and Goals

Role of Data Science at Space Y:

Compete with SpaceX using predictive models.

Key Objectives:

Launch Cost Prediction: Estimate costs using data analytics and machine learning.

First Stage Recovery Prediction: Predict Falcon 9 first stage landing success based on payload, orbit, and mission specifics.

Goals:

- Optimize launch costs.
- Enhance reliability and reusability of rocket stages.
- Improve Space Y's competitiveness in the commercial space industry.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  Data collected using the SpaceX API and the web scraping method

- Perform data wrangling

  In this data wrangling process, we cleaned and transformed launch data, focusing on attributes such as Flight Number, Date, Launch Site, Orbit, and Outcome. We standardized launch sites, categorized orbits, and converted landing outcomes into binary classification variables (0 for failure, 1 for success). This process enables efficient analysis of launch performance and outcomes, providing clear insights into the success rates of booster landings.Describe how data was processed

Methodology

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Compare Logistic Regression, SVM, Decision Trees, and K-Nearest Neighbors

Data Collection

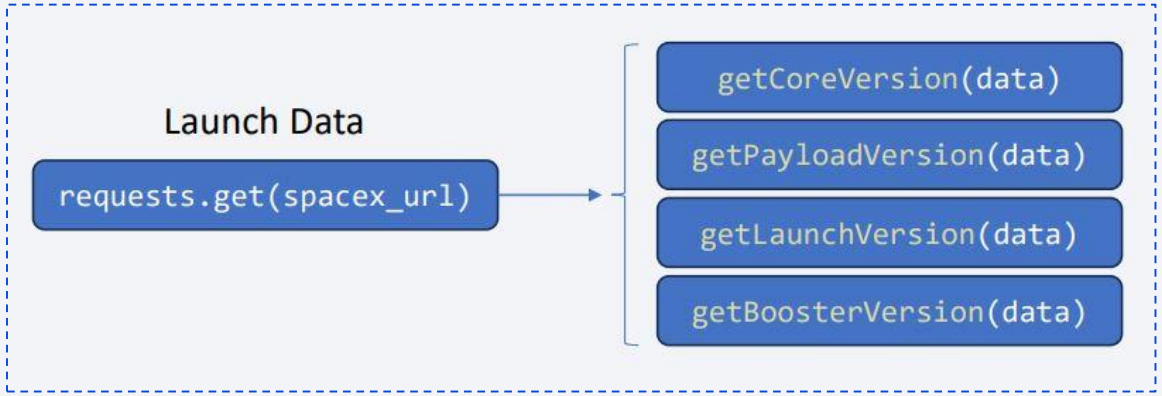
Steps:

1. Data Retrieval:
   - Source: SpaceX REST API (api.spacexdata.com/v4/launches/past)
   - Method: GET request, JSON response
   - Conversion: json_normalize to pandas dataframe
2. Data Enrichment:
   - Fetch detailed data using IDs for Booster, Launchpad, payload, core
   - Pre-created functions for specific data extraction
3. Data Cleaning:
   - Filter: Remove Falcon 1 launches
   - Null Handling: Replace PayloadMass nulls with mean
   - Column Handling: Leave LandingPad nulls for one-hot encoding
4. Web Scraping:
   - Tool: BeautifulSoup
   - Source: Related Wikipedia pages
   - Conversion: HTML tables to pandas dataframe

Data Collection – SpaceX API

To work with SpaceX's API for retrieving past launch data and organizing it into a dataframe, we can follow these steps:

- **Fetch Launch Data**: Use the SpaceX API to get past launch data.
- **Extract Launch IDs**: Extract the ID numbers from the fetched data.
- **Fetch Measured Data**: Use the IDs to fetch detailed measured data for each launch.
- **Generate Lists**: Store the data in lists.
- **Combine Lists into a DataFrame**: Combine these lists into a pandas dataframe for analysis or further processing.

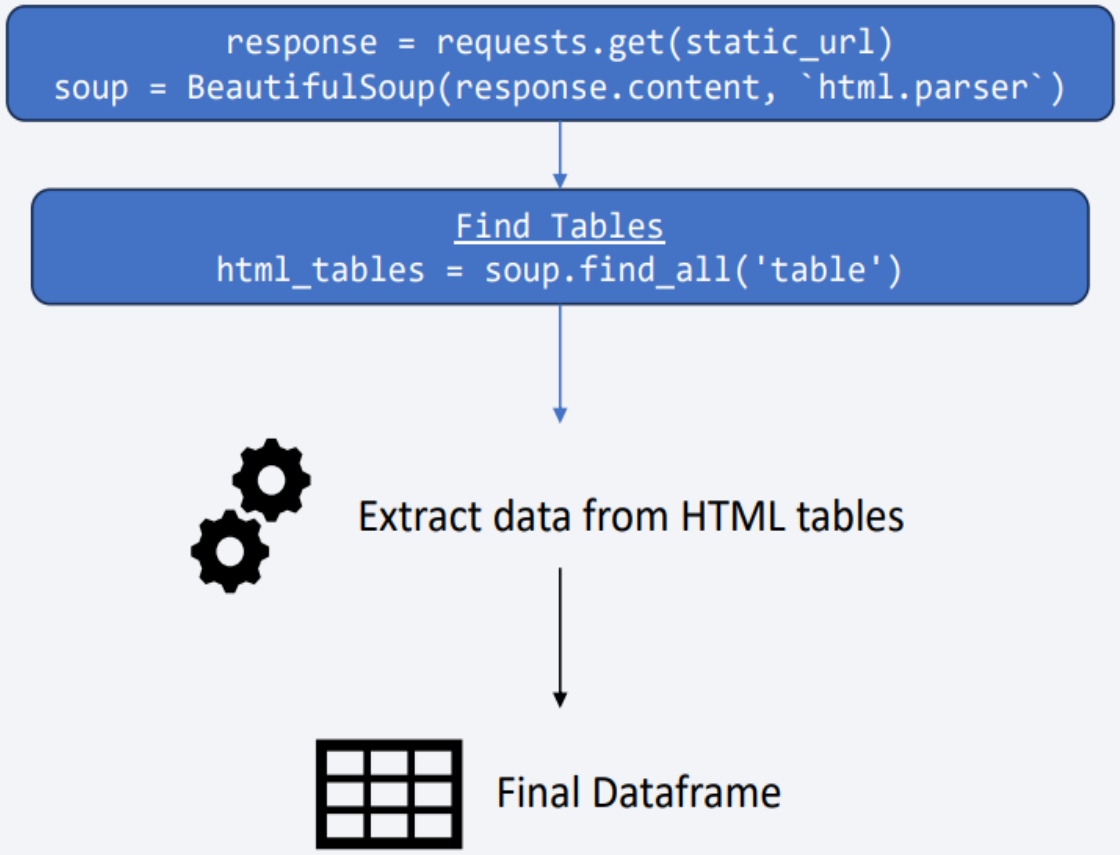


GitHub URL of the completed SpaceX API calls notebook

Data Collection - Scraping

- To web scrape Falcon 9 launch records from Wikipedia, used BeautifulSoup to extract the HTML table and then parse it into a Pandas dataframe.

[GitHub URL of the web scraping notebook]



Data Wrangling

1. **Load Data:** Load the combined dataset from both the API and Wikipedia.
2. **Inspect Data:** Understand the structure and check for missing values.
3. **Convert Outcomes to Labels:** Convert mission outcomes into binary labels for successful (1) and unsuccessful (0) landings.
4. **EDA:** Perform various visualizations to identify patterns and relationships in the data.

This process will prepare the data for training supervised models to predict SpaceX rocket landing outcomes.

GitHub URL of data wrangling notebooks

Landing Outcome	Good/Bad Launch?	Class
True ASDS	Good	1
None None	Bad	0
True RTLS	Good	1
False ASDS	Bad	0
True Ocean	Good	1
False Ocean	Bad	0
None ASDS	Bad	0
False RTLS	Bad	0

EDA with Data Visualization

- Scatter Plot for:
  1. FlightNumber vs. PayloadMass
  2. FlightNumber vs LaunchSite
  3. PayloadMass vs LaunchSite
  4. FlightNumber vs Orbit type
  5. Payload vs Orbit type
- Bar plot:
  1. Oribit vs Success Rate
- Line plot:
  1. launch success yearly trend

GitHub URL of EDA with data visualization notebook

EDA with SQL

- create table SPACEXTABLE as select * from SPACEXTBL where Date is not null
- select distinct("Launch_Site") from spacextable
- select "Launch_Site" from SPACEXTABLE where "Launch_Site" like "CCA%" limit 5
- select sum("PAYLOAD_MASS__KG_") from spacextable where "Customer" = 'NASA (CRS)'
- select avg("PAYLOAD_MASS__KG_") from spacextable where "Booster_Version" = 'F9 v1.1‘
- select "Date" from spacextable where "Landing_Outcome" like 'Success%' limit 1
- select distinct("Booster_Version") from spacextable where "Landing_Outcome" = 'Success (drone ship)' and "PAYLOAD_MASS__KG_" between 4000 and 6000
- select "Mission_Outcome",count("Mission_Outcome")as "Total" from spacextable group by "Mission_Outcome"
- select distinct("Booster_Version") from spacextable where "PAYLOAD_MASS__KG_" = (select max("PAYLOAD_MASS__KG_") from spacextable)
- select substr(Date, 6,2) as "Month","Landing_Outcome","Booster_Version","Launch_Site" from spacextable where "Date" like '2015%' and "Landing_Outcome" = 'Failure (drone ship)'
- select "Landing_Outcome",count("Landing_Outcome")as "Total" from spacextable where "Date" between '2010-06-04' and '2017-03-20' group by "Landing_Outcome"

GitHub URL of EDA with SQL notebook

Build an Interactive Map with Folium

- A global map was generated to identify any commonalities in location between the launch sites.
  - Each site was labeled with a marker and pop-up label
  - Sites are clustered into Marker Cluster objects for a more appealing visualization
  - Launch Outcome Disposition was color-coded for each site

GitHub URL of interactive map with Folium map

Build a Dashboard with Plotly Dash

Pie Chart for summarizing dispositions:

- Proportion of successful launches between all Launch Sites
- Success and fail rate for each site

Scatter Plot to identify any trends or success indicators:

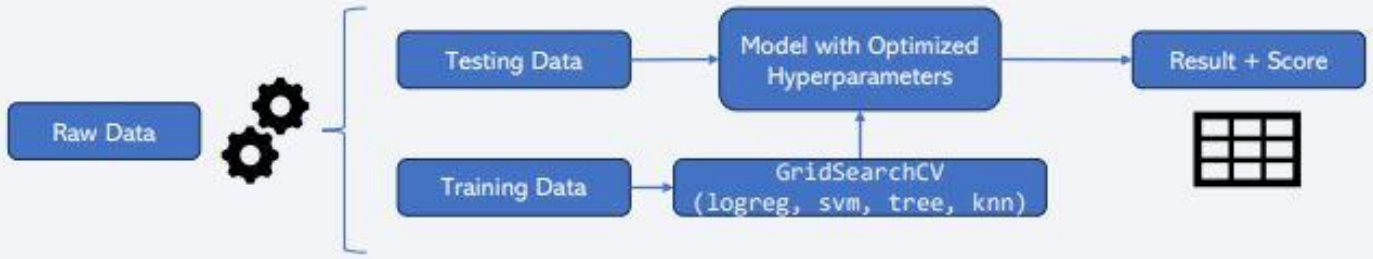
- Launch Outcome Disposition by Payload Mass, Booster Version, and Launch Site

- GitHub URL completed Plotly Dash lab

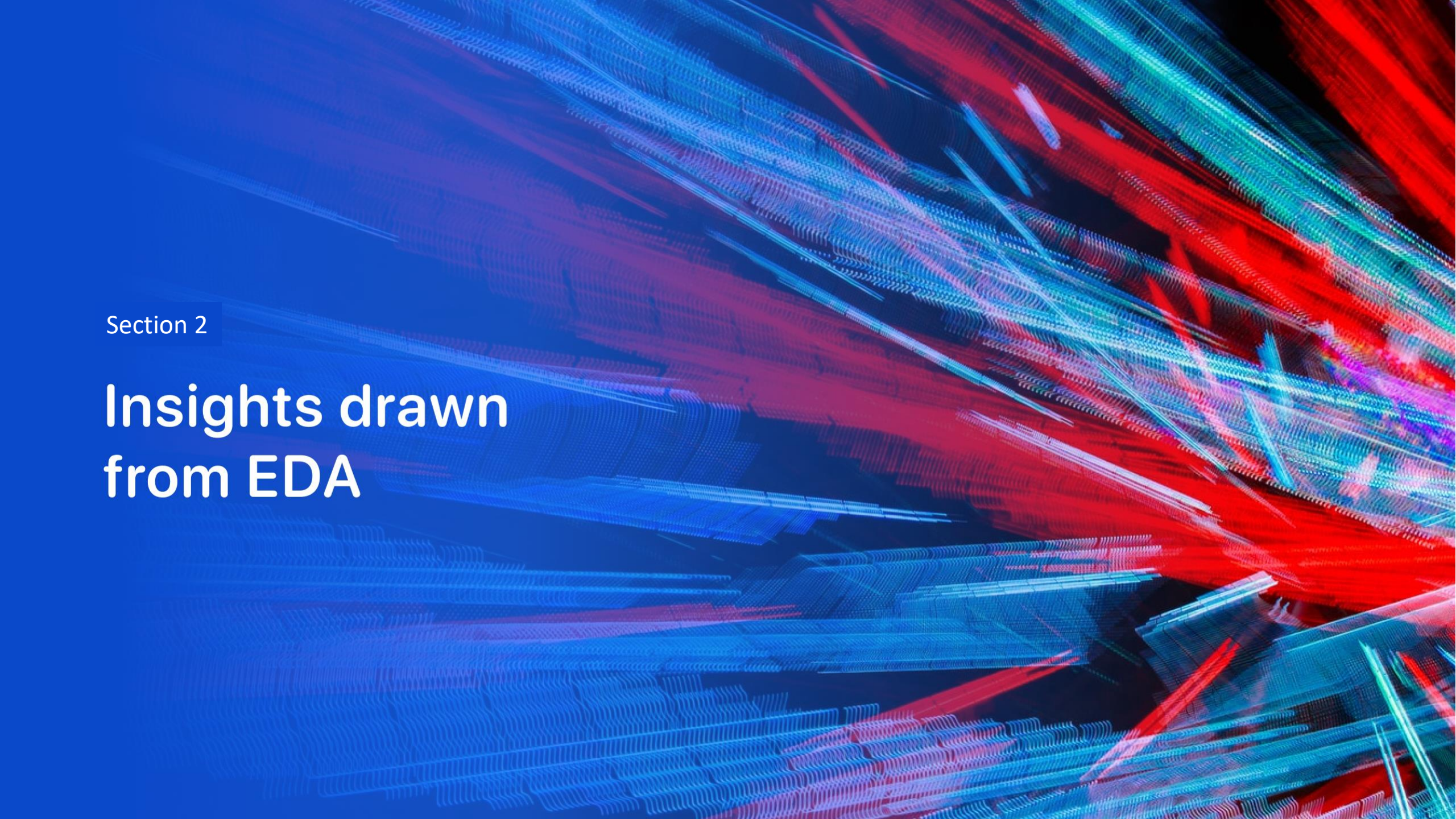
Predictive Analysis (Classification)

Compared 4 different machine learning classification algorithms

- Logistic Regression, SVM, Decision Tree, and K-Nearest Neighbors
- Data is preprocessed using one-hot encoding and standardization
- Data was split into training and testing sets using sklearn train_test_split
- Models were trained and hyperparameters were optimized using sklearn GridSearchCV



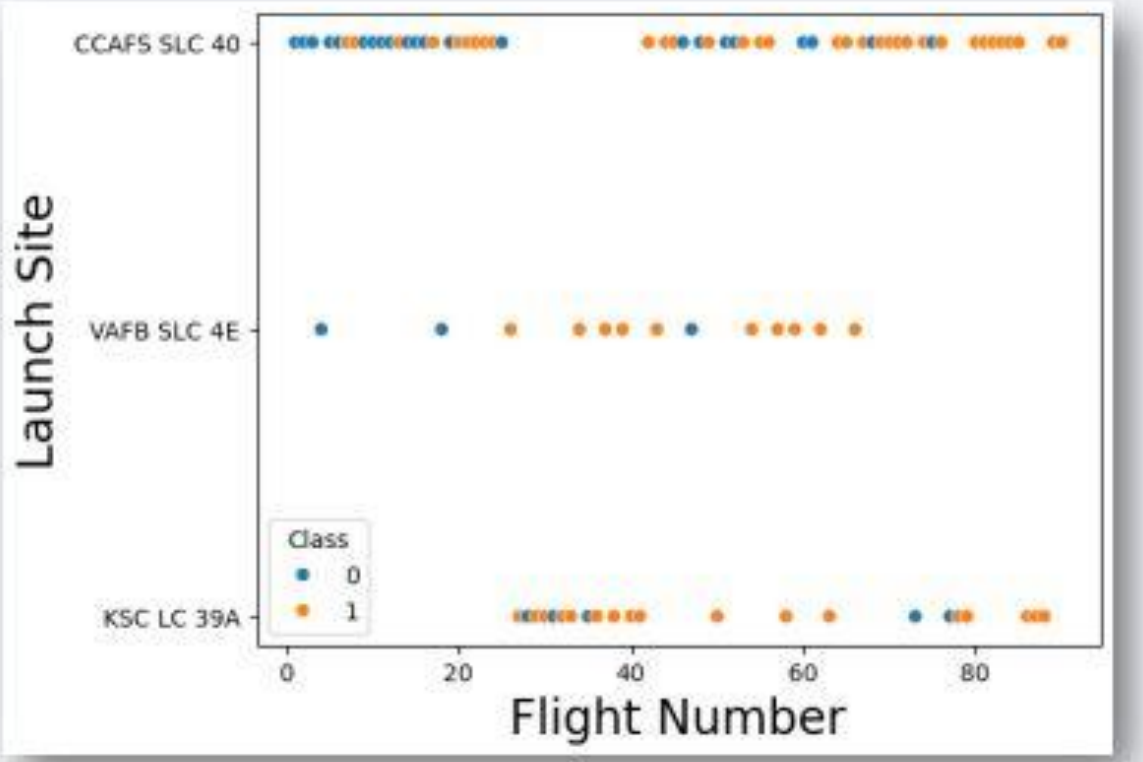
GitHub URL predictive analysis lab


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

Observations:

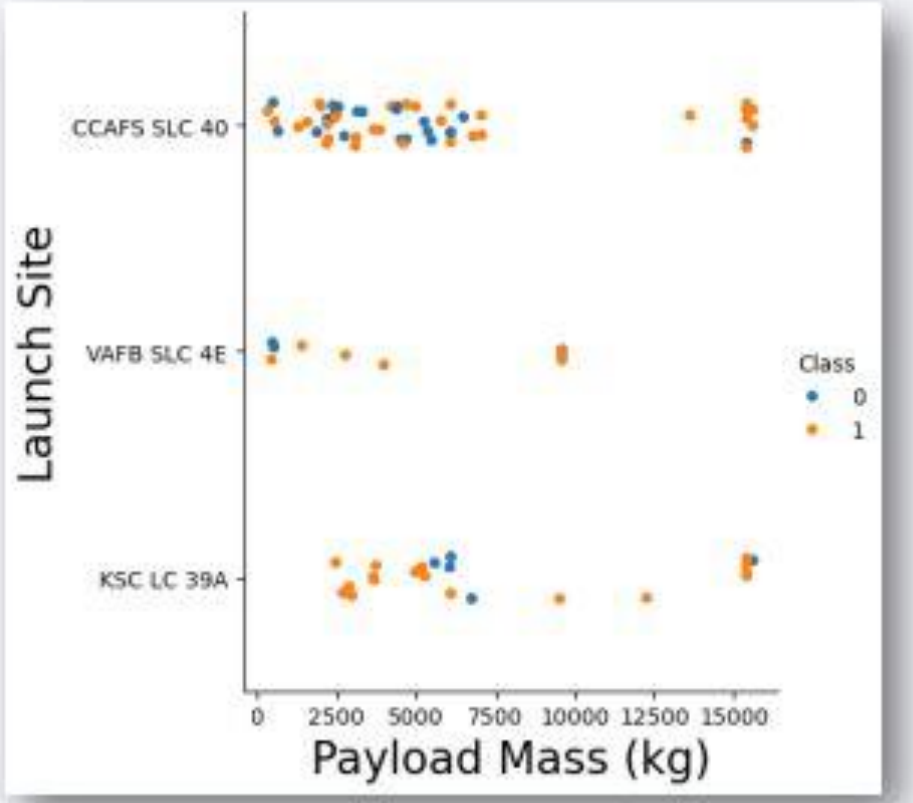
- Earlier flights have higher rate of failure
- Earlier flights are mostly at CCAFS SLC 40.



Payload vs. Launch Site

Observations:

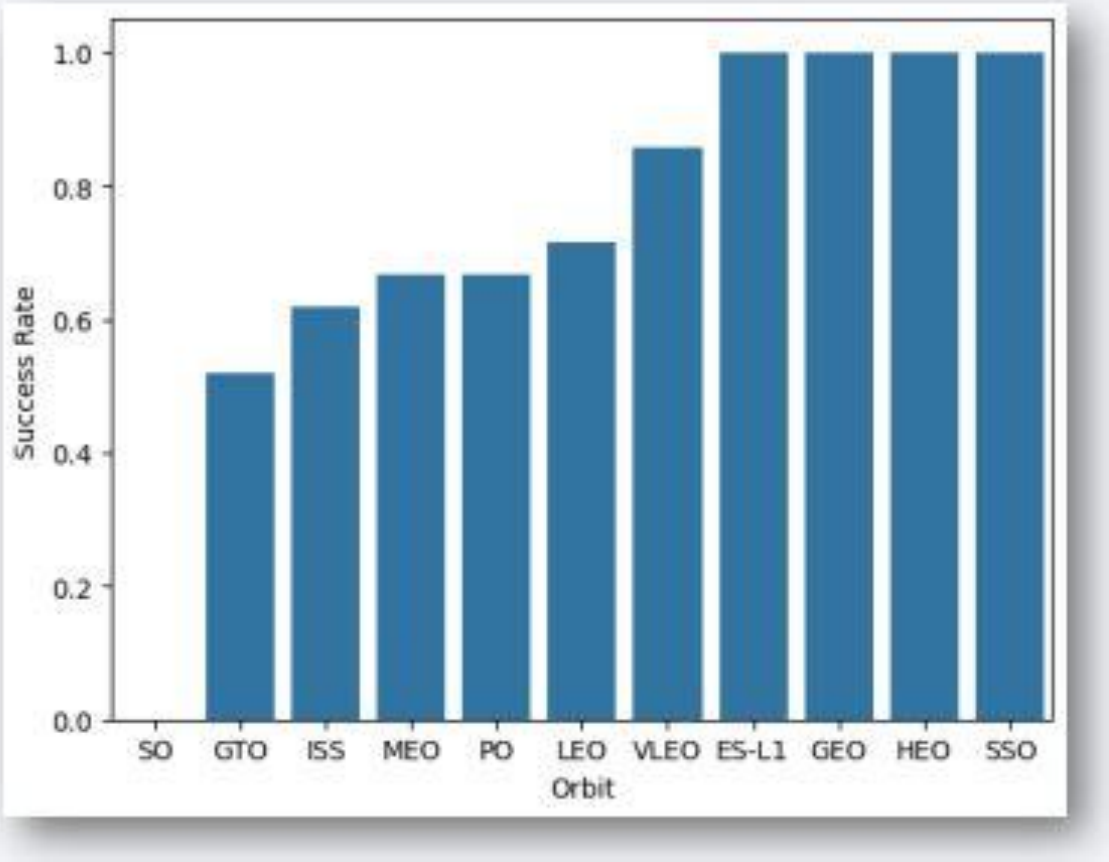
- High payload mass >7500 has a high success rate
- There is a boundary that can be inferred at 6000 kg for launch site KSC LC 39A.



Success Rate vs. Orbit Type

Observations:

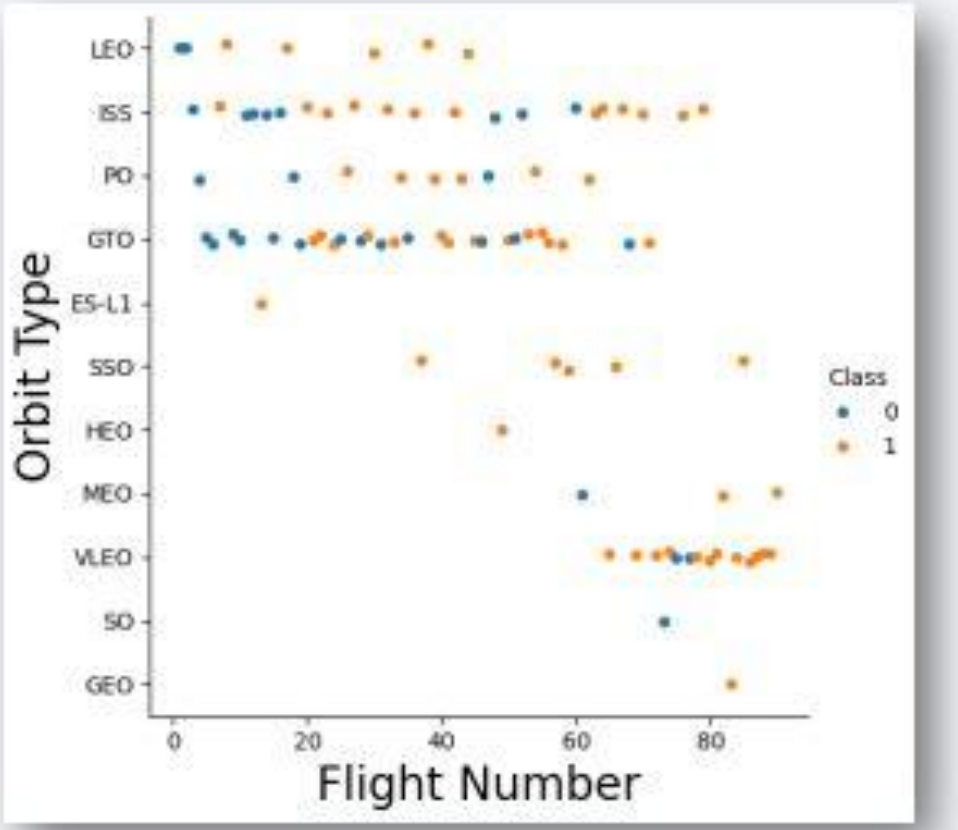
- ES-L1, GEO, HEO, and SSO Orbit types have 100% landing success rate.
- SO has a 0% success rate.



Flight Number vs. Orbit Type

Observations:

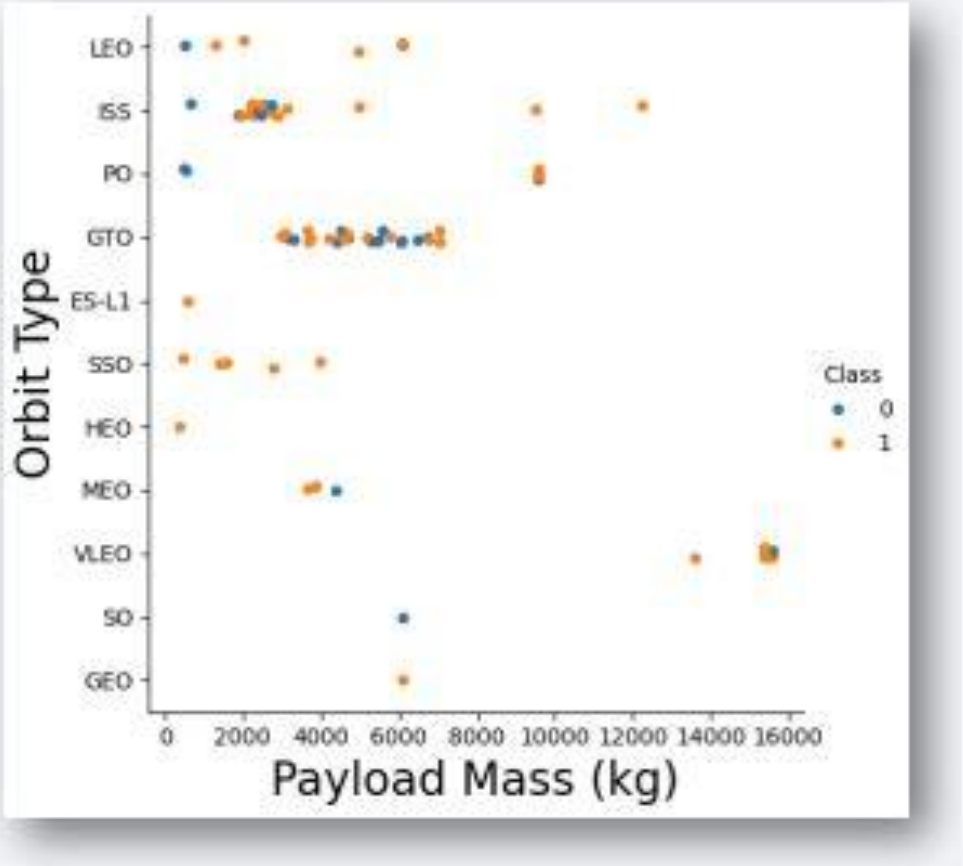
- ES-L1, GEO, HEO, and SSO Orbit types have 100% landing success rate, **however** the launch sample size is small.
- The Orbit types with the most frequent launches are ISS, GTO, and VLEO.
- VLEO has the highest success rate after excluding orbits with 100% and 0% success rates.



Payload vs. Orbit Type

Observations:

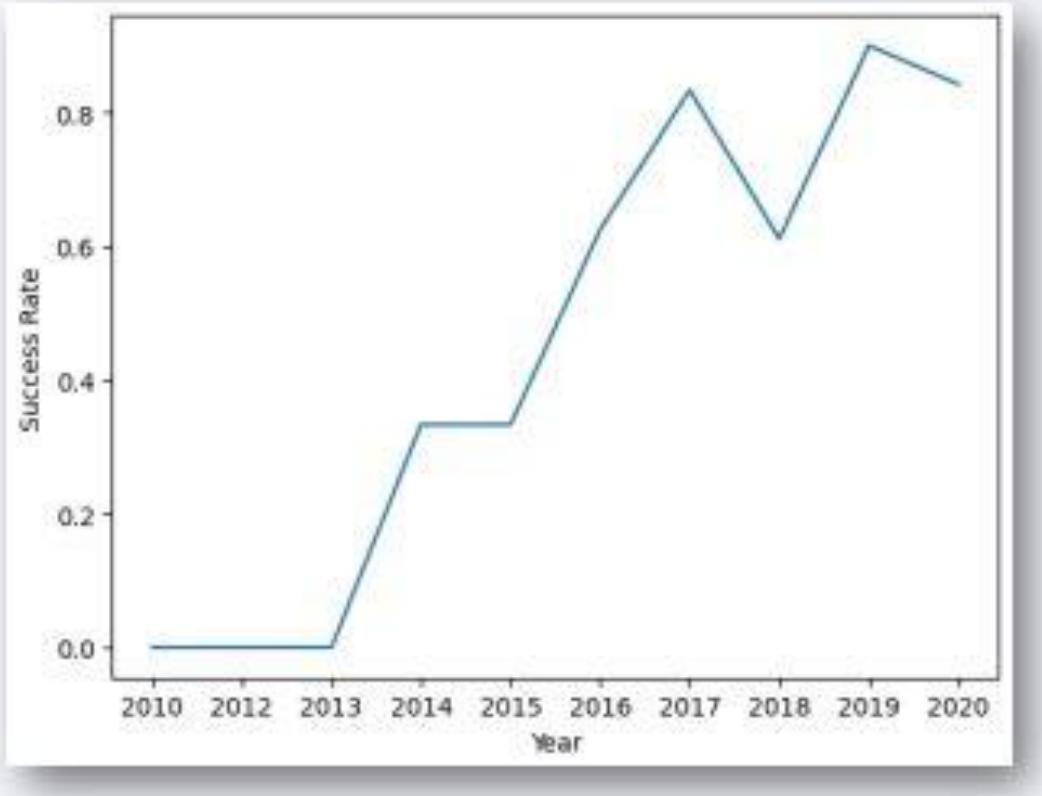
- Certain orbit types have specific payload mass ranges



Launch Success Yearly Trend

Observations:

- Launches improve over time, as illustrated by the launch success increase over the years.



All Launch Site Names

- Unique launch sites can be found by applying the DISTINCT function on the launch site column.
- The different launch sites are CCSFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40

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

Launch Site Names Begin with 'CCA'

- Displaying 5 full records where launch site names start with `CCA`.

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 from NASA (CRS) is 48,213 kg.

SUM(PAYLOAD_MASS__KG_)
48213

- The mass of all payloads where the customer name contains NASA and CRS was summed together.

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is 2928.4 kg.
- The average of all payload masses where the Booster Version contains F9 and v1.1 was calculated.

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- The minimum date, 2015-12-22, was selected from records where the landing outcome contains `success` and `ground pad`

Min(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- List the names of 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 unique booster versions that have a successful landing outcome on a drone ship with payload mass between 4000 and 6000 are F9 FT B1022, F9 FT B1026, F9 FtB1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- Count all mission outcomes, grouping by the mission outcome.

Mission_Outcome	COUNT(Mission_Outcome)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- The boosters versions that have carried the maximum payload mass are shown in the table.

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

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

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

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

- 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, in descending order:

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

Section 3

Launch Sites Proximities Analysis

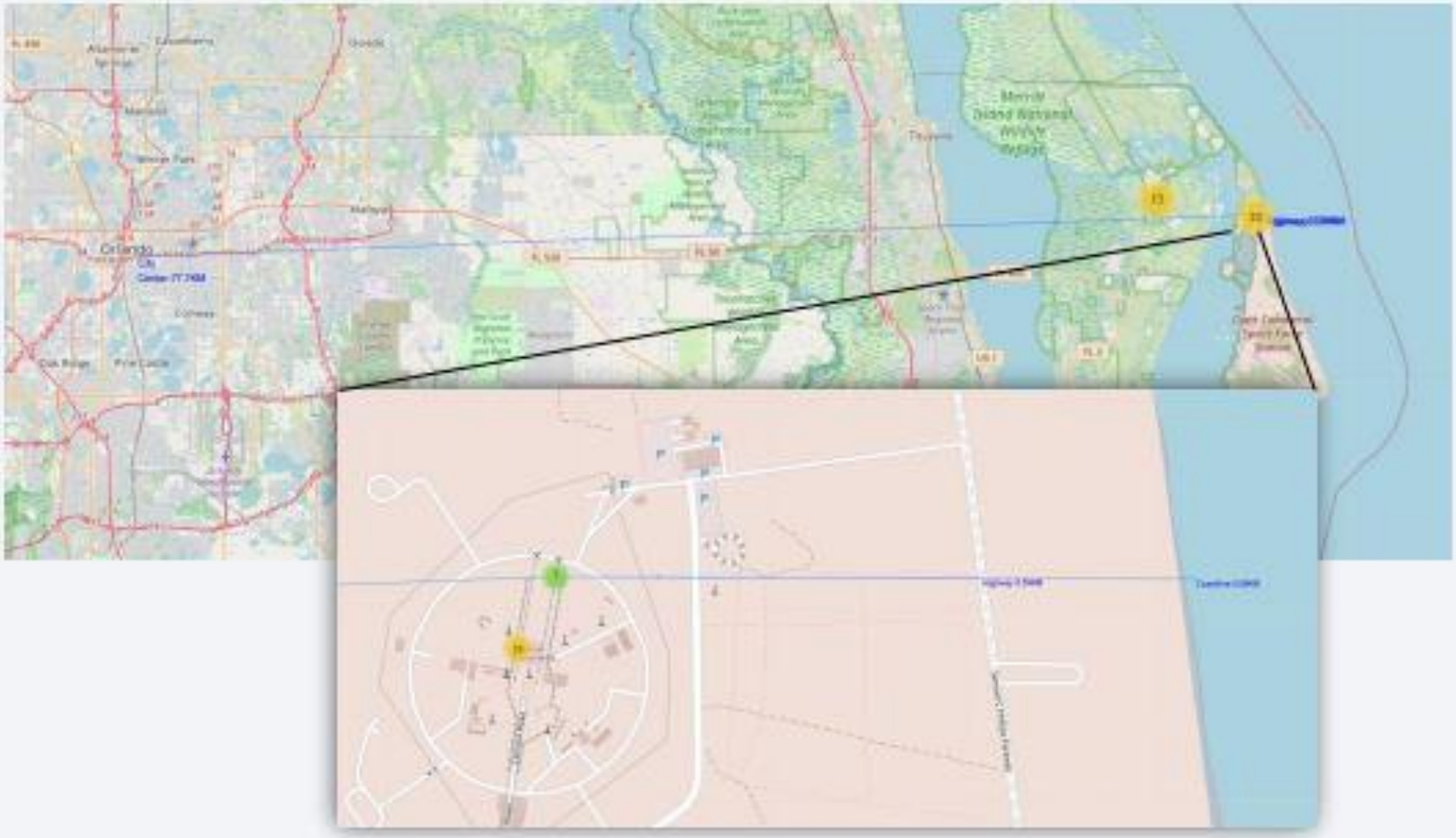
SpaceX Launch Site Locations

- 3 launch sites are in close proximity to each other in Florida
- Last site is on the West Coast in Southern California

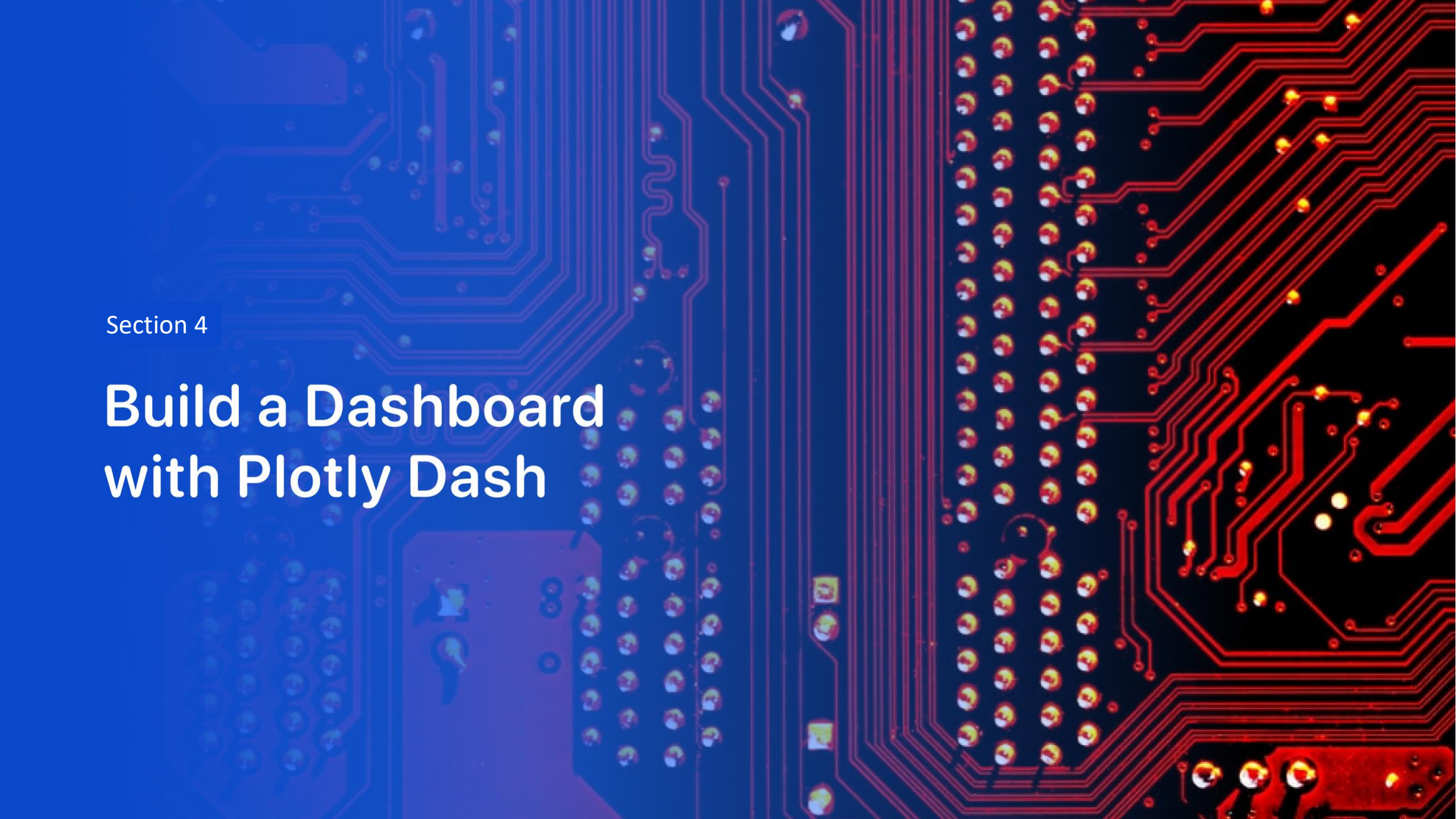
Launch Outcomes by Site

- Showing launch outcomes for launch site CCAFS SLC-40
- Green indicates success, red indicates fail

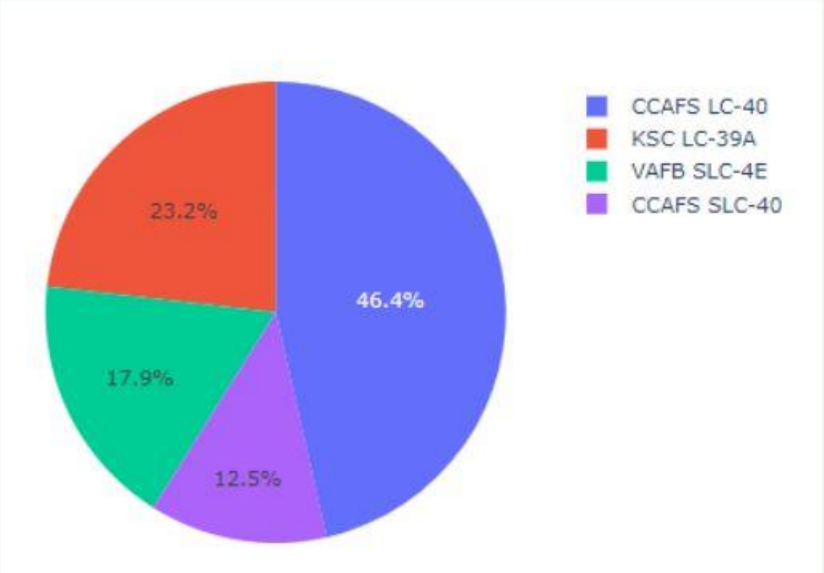
Proximity Analysis



- Showing proximity to nearest coast, highway, and city for launch site CCAFS SLC-40
- Launch sites are built near to the coast and a highway, located far from highly populated cities.


Section 4
Build a Dashboard with Plotly Dash

Launch Success Dashboard

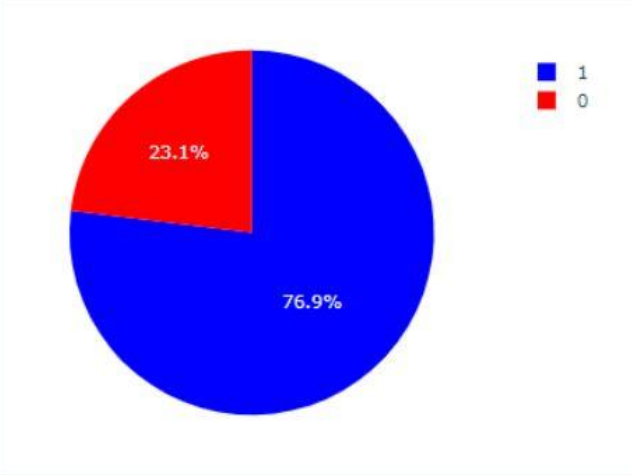


The percentage breakdown of successful launches by site is illustrated by the pie chart.

CCAFS LC-40 has the highest proportion out of all successful launches, followed by KSC LC-39A, VAFB SLC-4E, and CCAFS SLC-40

Launch Site-wise Success Rate

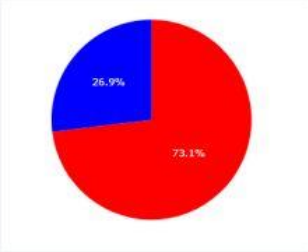
KSC LC-39A



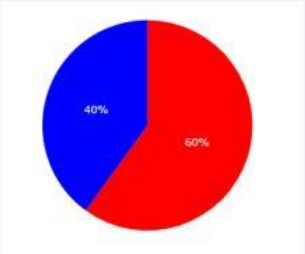
The site with the highest success rate is KSC LC-39A, with 76.9% of their launches being successful.

This is over 30% greater than the next highest success rate of 42.9%.

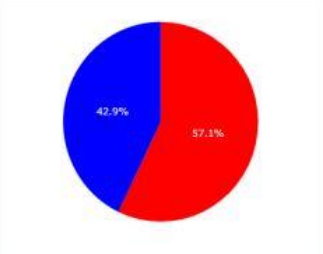
CCAFS LC-40



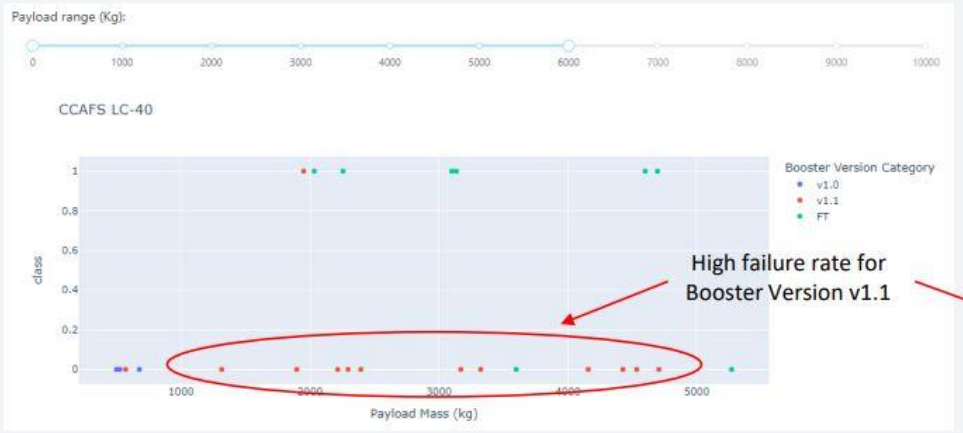
VAFB SLC-4E

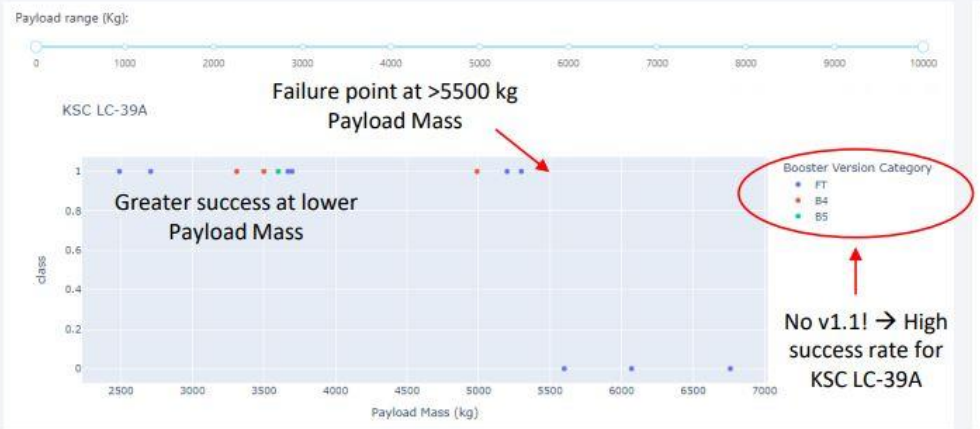


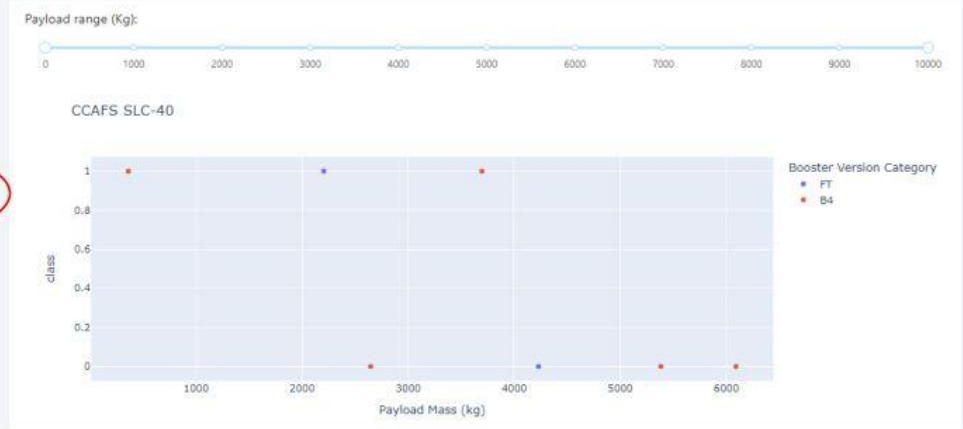
CCAFS SLC-40



Launch Outcome Dashboard





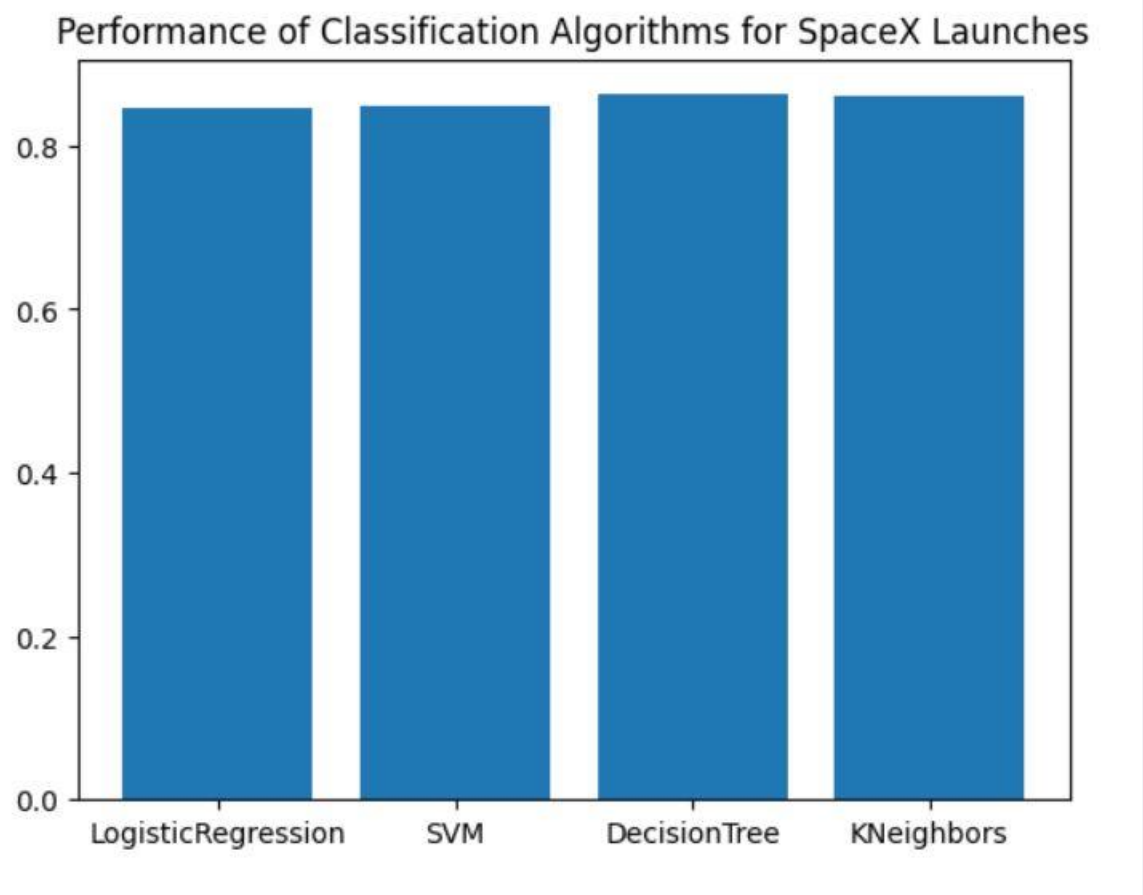


Section 5

Predictive Analysis (Classification)

Classification Accuracy

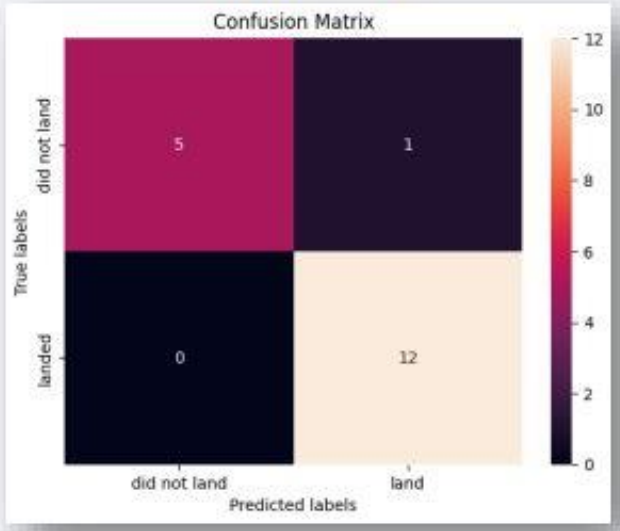
- Decision Tree model has the highest classification accuracy



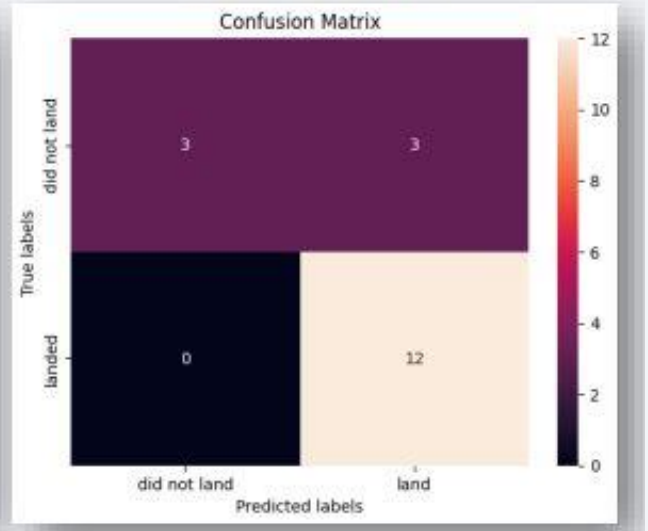
Confusion Matrix

- Tree algorithm provides the best accuracy with the given dataset
- From confusion matrix, only 1 landing was misclassified as landed when it failed
- Other algorithms misclassified 3 failed landings as successes

Decision Tree



Other Algorithms



Conclusions

1. Location Advantages

- Launch sites are typically located near highways and ocean coasts.
- These locations allow for quicker supply and transport.
- Coastal sites serve as gateways to Asia (Pacific) and Europe (Atlantic).

2. Success Rates

- The success rate at Kennedy Space Center (KSC) Launch Complex 39A is more than 30% better than the next best site.
- The higher success rate might be influenced by the absence of launches of the low-success-rate Booster Version v1.1 from this site.

3. Annual Success Rate Trends

- The success rate generally increases year over year, with the exception of 2018.

4. Prediction Model

- A Decision Tree classification model correctly classifies 94.4% of the training data launches.
- This makes it the strongest performing prediction model for the provided data.

Appendix

Thank you!