APPLIED DATA SCIENCE CAPSTONE PROJECT

SpaceX Falcon 9 Landing Analysis

- MD Humayun Kabir

OUTLINE

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

EXECUTIVE SUMMARY

This project analyzes SpaceX launch data to predict whether the Falcon 9 first stage will land successfully. Key steps include:

1. Data Preparation – Cleaning and formatting launch records.

2. Exploratory Analysis – Identifying trends and influential factors.

3. Interactive Visualizations – Exploring correlations with success rates.

4. Machine Learning – Training models, with Decision Trees performing best.

The results help understand which launch conditions lead to successful landings, demonstrating the value of ML in aerospace.

INTRODUCTION

- SpaceX's Falcon 9 offers cost-effective launches at 62 million —much cheaper than competitors 165 million — thanks to reusable first stages. By predicting whether the first stage will land successfully, we can estimate launch costs, providing valuable insights for competitive bidding.
- While some ocean landings are intentional, our goal is to determine landing success based on key factors like payload mass, orbit type, and launch site. Using machine learning, we'll analyze these features to answer: Will the Falcon 9's first stage land successfully?
- This project will help assess SpaceX's cost advantage and support strategic decision-making in the space launch industry.

METHODOLOGY

The overall methodology includes:

1. Data collection, wrangling, and formatting, using:
   - SpaceX API
   - Web scraping
2. Exploratory data analysis (EDA), using:
   - Pandas and NumPy
   - SQL
3. Data visualization, using:
   - Matplotlib and Seaborn
   - Folium
   - Dash
4. Machine learning prediction, using
   - Logistic regression
   - Support vector machine (SVM)
   - Decision tree
   - K-nearest neighbors (KNN)

METHODOLOGY : Data collection, wrangling, and formatting

SpaceX API:

We obtained launch data from SpaceX's official API (https://api.spacexdata.com/v4/rockets/), focusing exclusively on Falcon 9 missions.

Data Cleaning Process:

- Filtered to include only Falcon 9 launches
- Handled missing values by imputing column means
- Final dataset contains 90 launches with 17 key features

The processed dataset captures critical launch parameters needed for our landing success prediction analysis. Below is a preview of the structured data:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857
5	6	2014-01-06	Falcon 9	3325.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1005	-80.577366	28.561857

METHODOLOGY : Data collection, wrangling, and formatting

Web Scrapping:

- The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The website contains only the data about Falcon 9 launches.
- We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	F9 v1.07B0003.18	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.07B0004.18	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.07B0005.18	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success	F9 v1.07B0006.18	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success	F9 v1.07B0007.18	No attempt\n	1 March 2013	15:10

METHODOLOGY : Exploratory Data Analysis (EDA)

- Pandas and NumPy
  - Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:
    - The number of launches on each launch site
    - The number of occurrence of each orbit
    - The number and occurrence of each mission outcome
- SQL
  - The data is queried using SQL to answer several questions about the data such as:
    - The names of the unique launch sites in the space mission
    - The total payload mass carried by boosters launched by NASA (CRS)
    - The average payload mass carried by booster version F9 v1.1

METHODOLOGY : Data Visualization

- Matplotlib and Seaborn
  - Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.
  - The plots and charts are used to understand more about the relationships between several features, such as:
    - The relationship between flight number and launch site
    - The relationship between payload mass and launch site
    - The relationship between success rate and orbit type
- Folium
  - Functions from the Folium libraries are used to visualize the data through interactive maps.
  - The Folium library is used to:
    - Mark all launch sites on a map
    - Mark the succeeded launches and failed launches for each site on the map
    - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

METHODOLOGY : Data Visualization

- Dash
  - Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
  - Using a pie chart and a scatterplot, the interactive site shows:
    - The total success launches from each launch site
    - The correlation between payload mass and mission outcome (success or failure) for each launch site

METHODOLOGY : Machine Learning Prediction

- Functions from the Scikit-learn library are used to create our machine learning models.
- The machine learning prediction phase include the following steps:
  - Standardizing the data
  - Splitting the data into training and test data
  - Creating machine learning models, which include:
    - Logistic regression
    - Support vector machine (SVM)
    - Decision tree
    - K nearest neighbors (KNN)
  - Fit the models on the training set
  - Find the best combination of hyperparameters for each model
  - Evaluate the models based on their accuracy scores and confusion matrix

RESULTS

The results are split into 5 sections:

- SQL (EDA with SQL)
- Matplotlib and Seaborn (EDA with Visualization)
- Folium
- Dash
- Predictive Analysis

In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

RESULTS : SQL – EDA with SQL

- The names of the unique launch sites in the space mission

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

- 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
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

RESULTS : SQL – EDA with SQL

- The total payload mass carried by boosters launched by NASA (CRS)

Total_Payload
45596

- The average payload mass carried by booster version F9 v1.1

AVG_Payload
2928.4

- The date when the first successful landing outcome in ground pad was achieved

First_Successfull_Landing
2015-12-22

RESULTS : SQL – EDA with SQL

- The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Booster_Version	PAYLOAD_MASS__KG_
F9 FT B1022	4696
F9 FT B1026	4600
F9 FT B1021.2	5300
F9 FT B1031.2	5200

- The total number of successful and failure mission outcomes

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

RESULTS : SQL – EDA with SQL

- The names of the booster versions which have carried the maximum payload mass

Booster_Version	Max_Payload
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

RESULTS : SQL – EDA with SQL

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

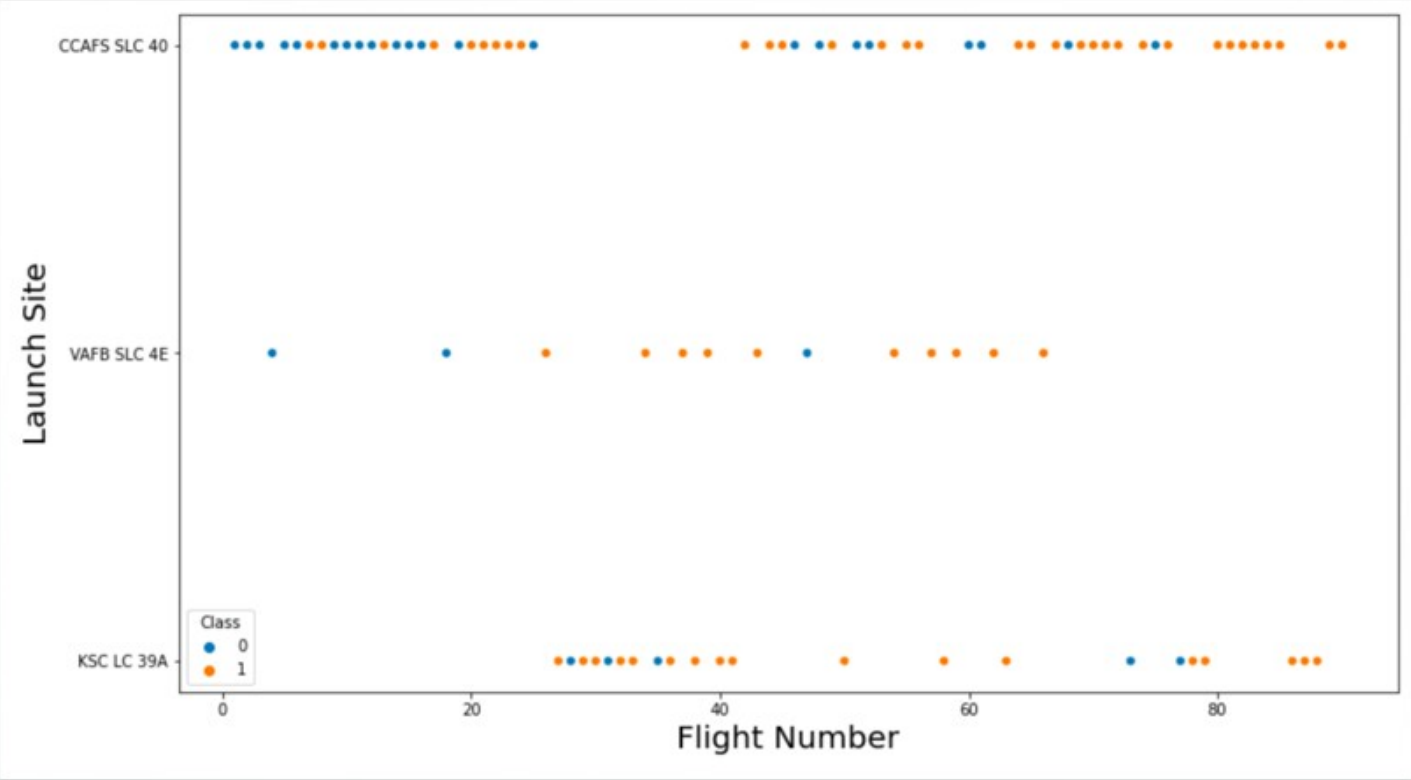
Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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

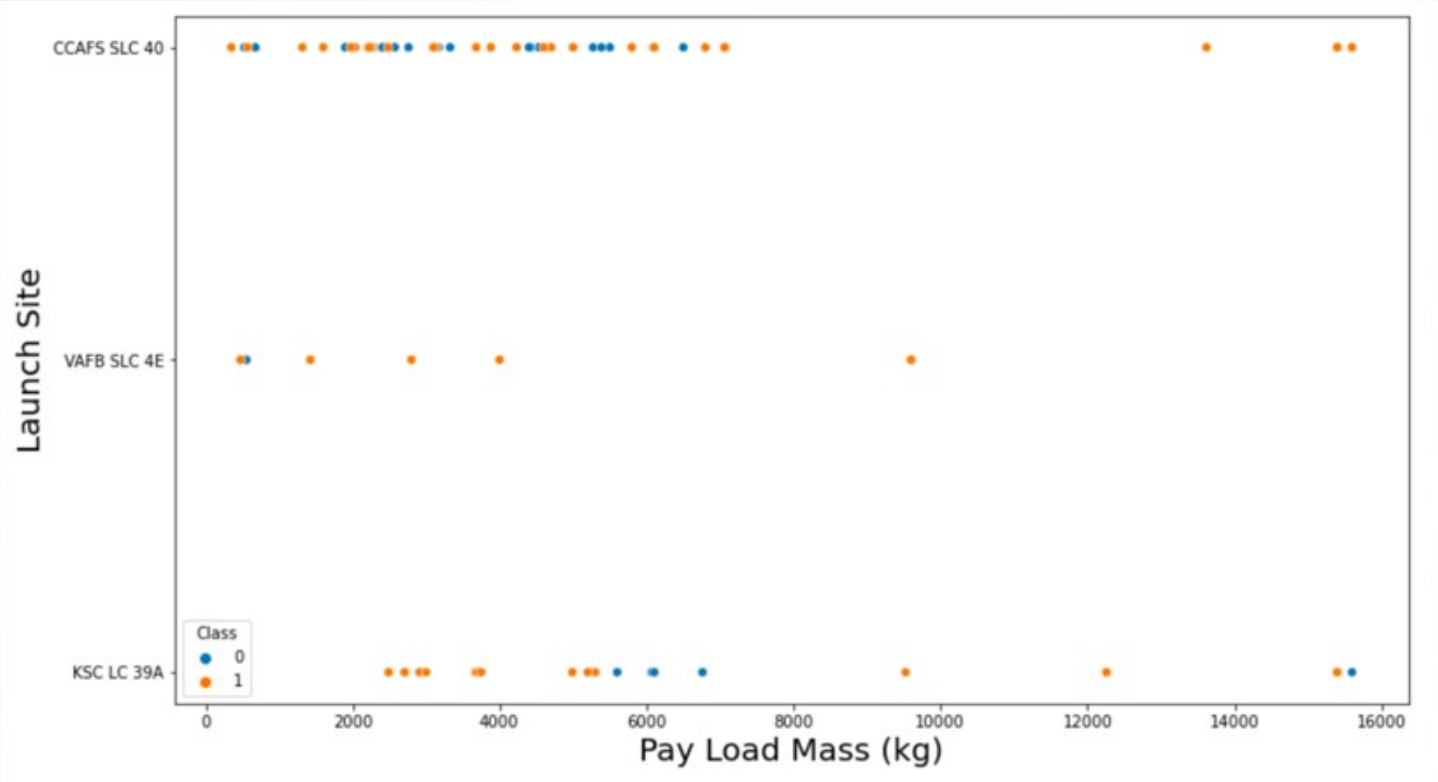
RESULTS : Matplotlib and Seaborn (EDA with Visualization)

- The relationship between flight number and launch site



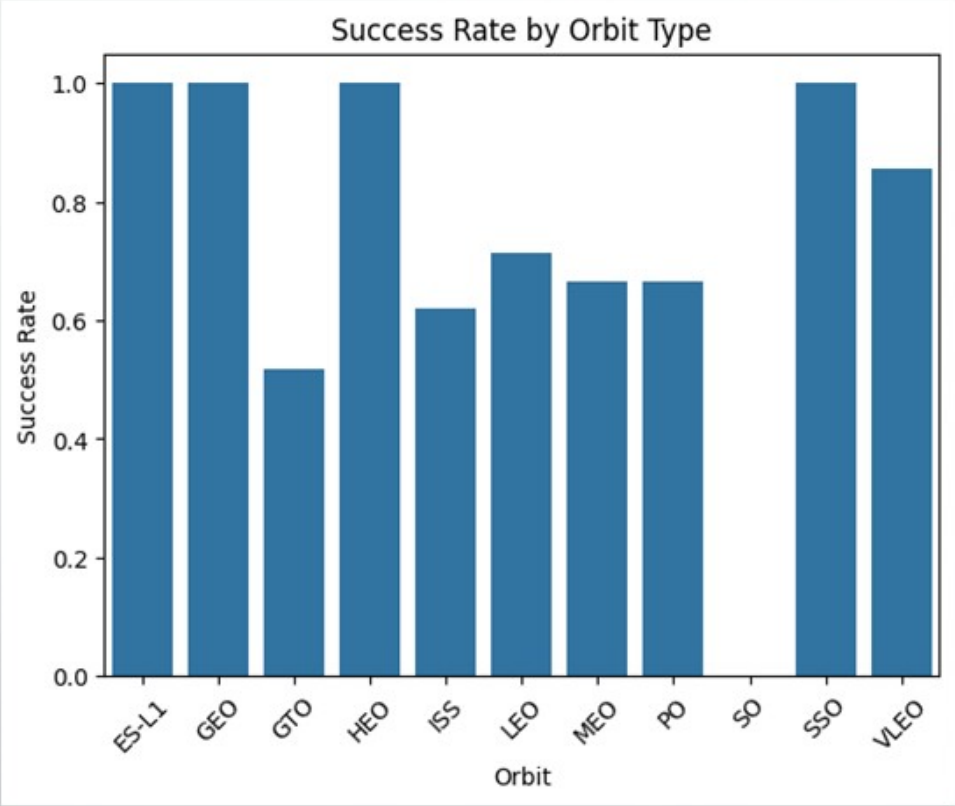
RESULTS : Matplotlib and Seaborn (EDA with Visualization)

- The relationship between payload mass and launch site



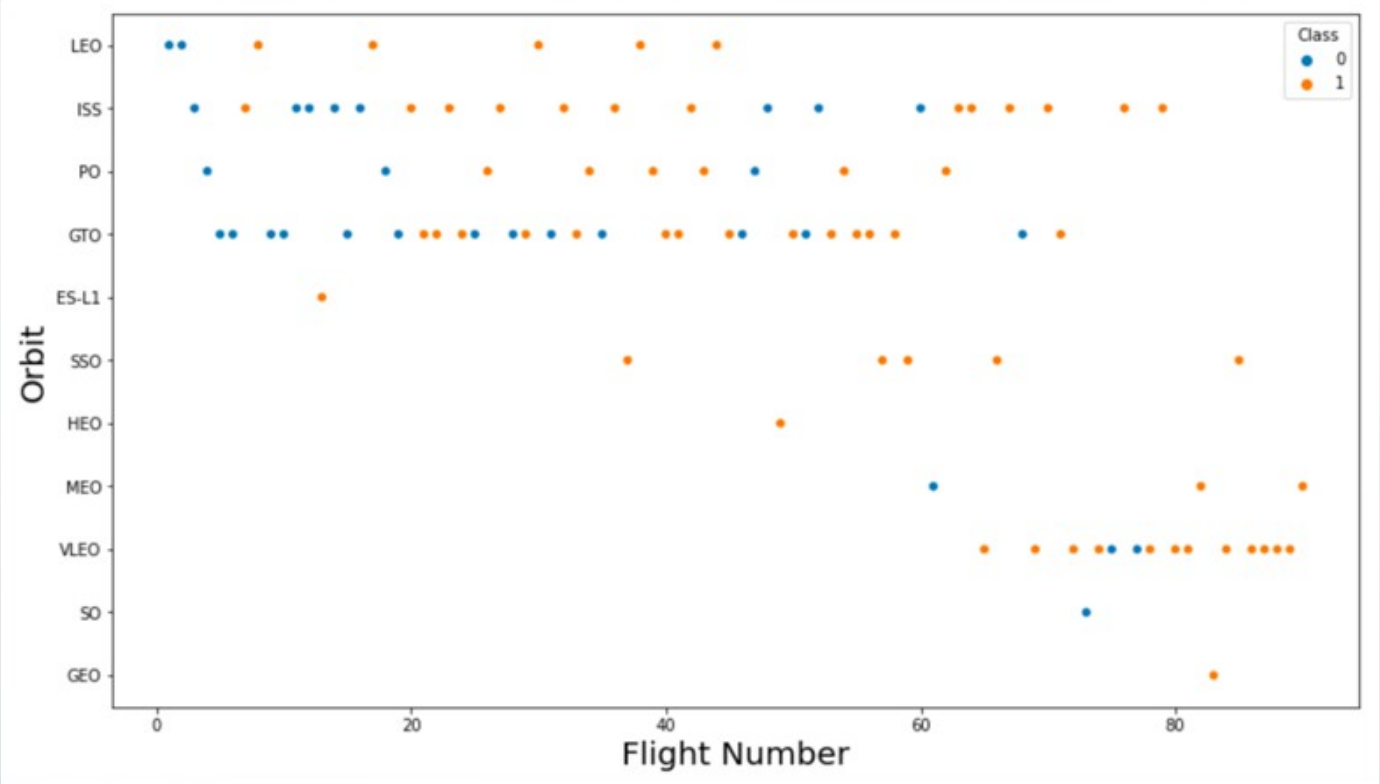
RESULTS : Matplotlib and Seaborn (EDA with Visualization)

- The relationship between success rate and orbit type



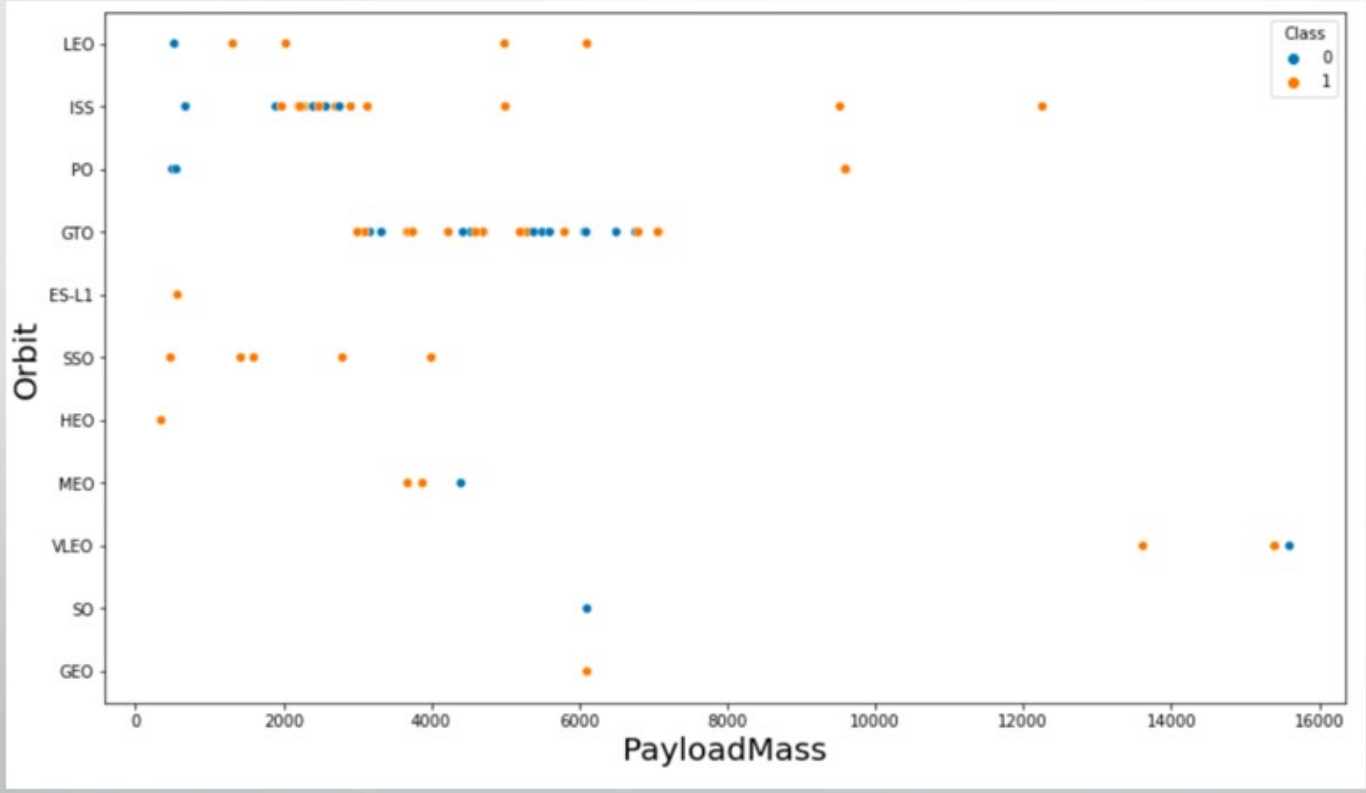
RESULTS : Matplotlib and Seaborn (EDA with Visualization)

- The relationship between flight number and orbit type



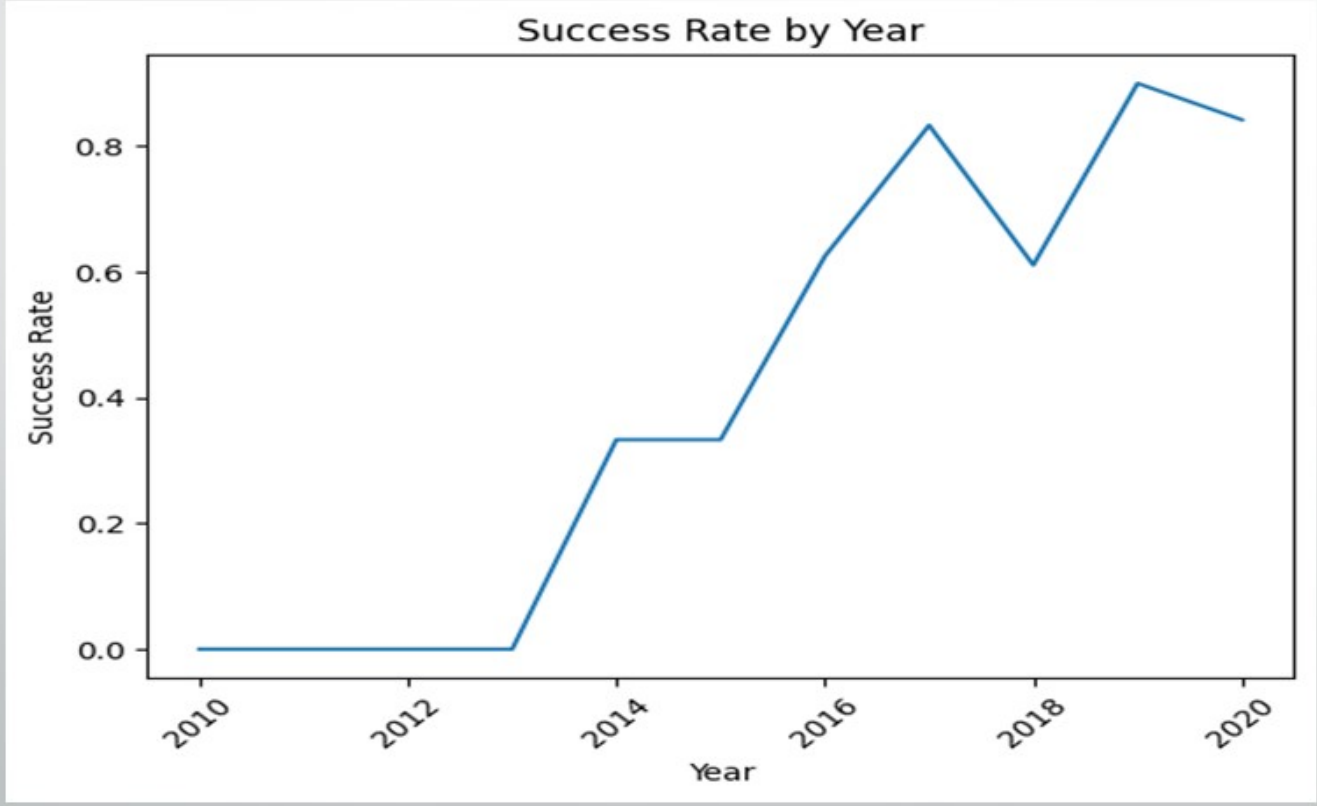
RESULTS : Matplotlib and Seaborn (EDA with Visualization)

- The relationship between payload mass and orbit type



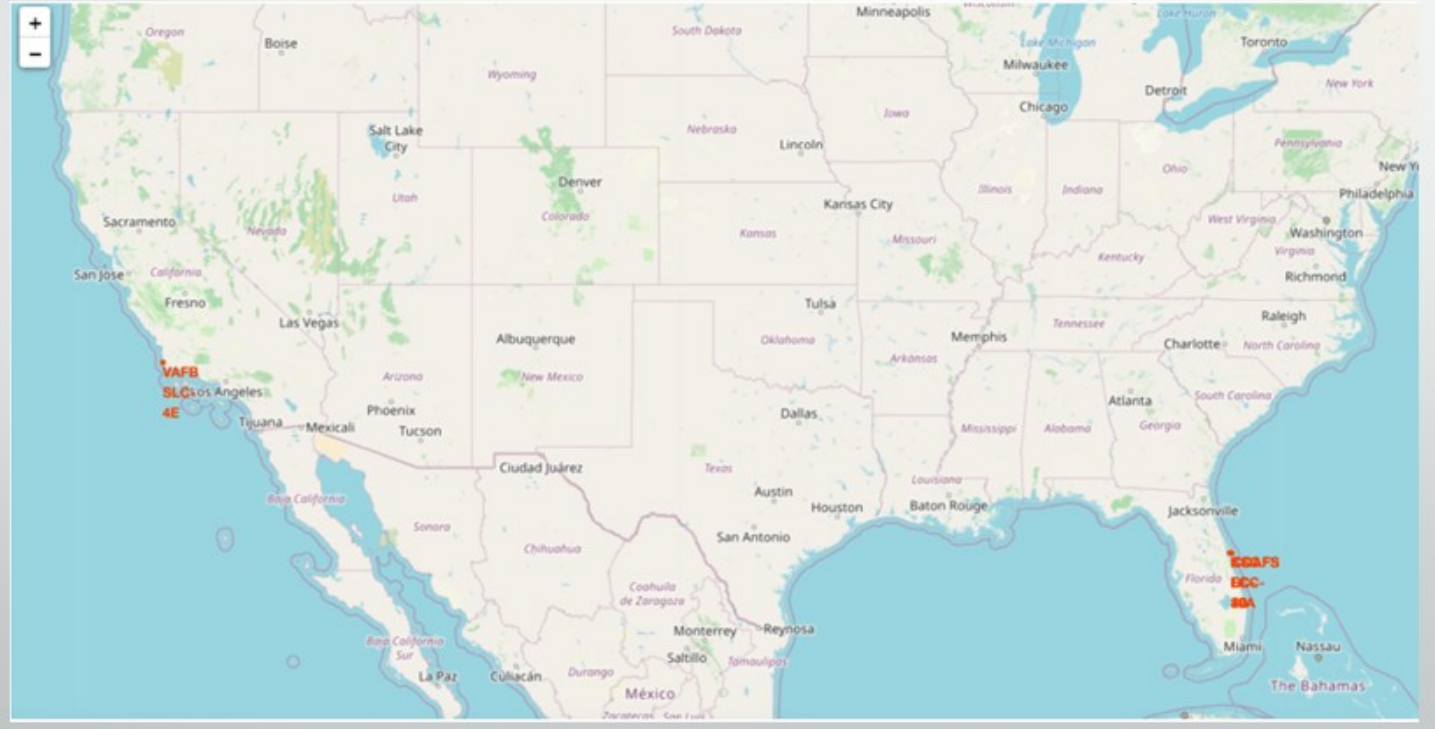
RESULTS : Matplotlib and Seaborn (EDA with Visualization)

- The launch success yearly trend



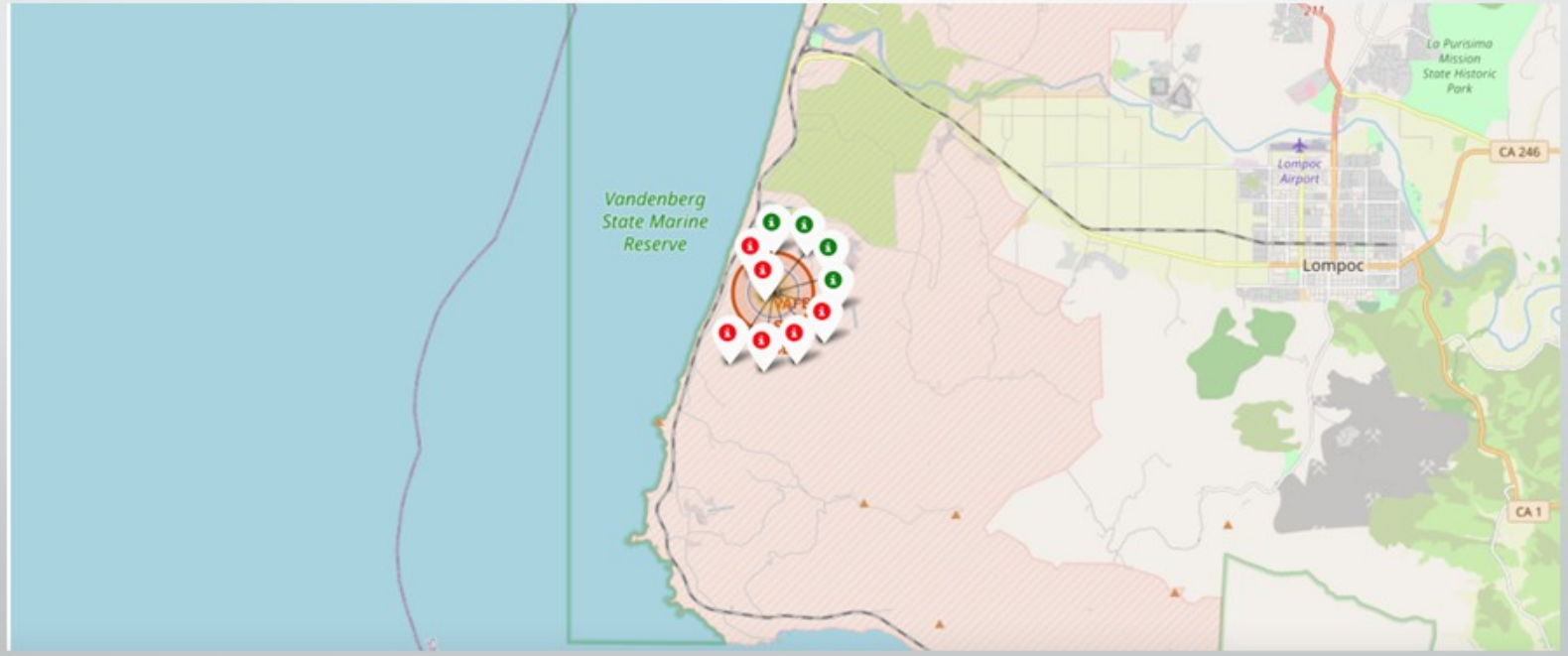
RESULTS : Folium

- All launch sites on map



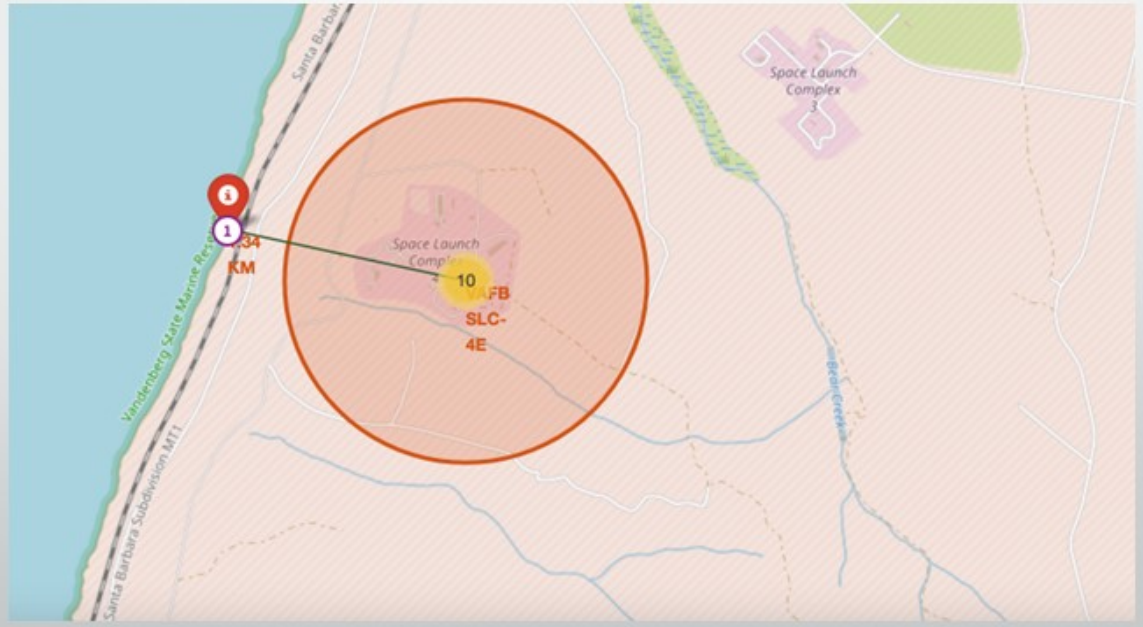
RESULTS : Folium

- The succeeded launches and failed launches for each site on map
  - If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch



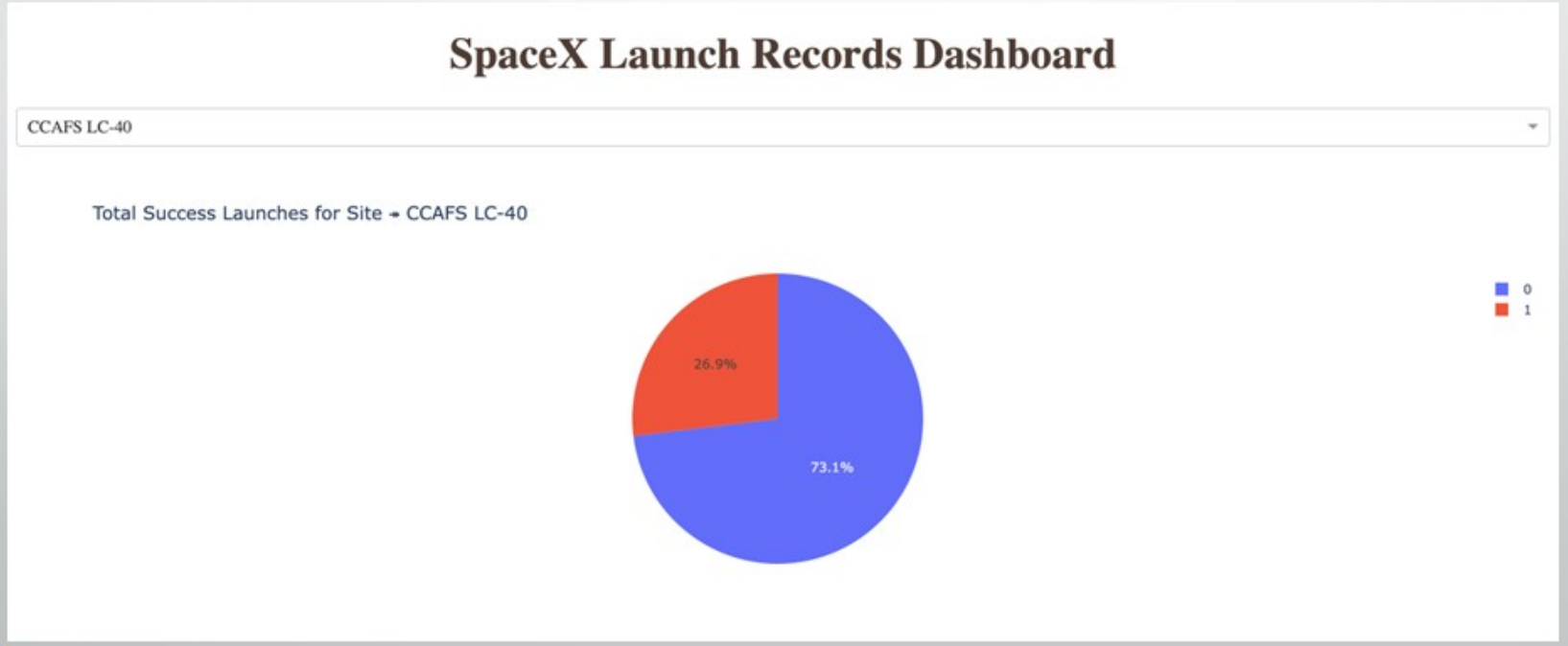
RESULTS : Folium

- The distances between a launch site to its proximities such as the nearest city, railway, or highway
- The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



RESULTS : Dash

- The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.

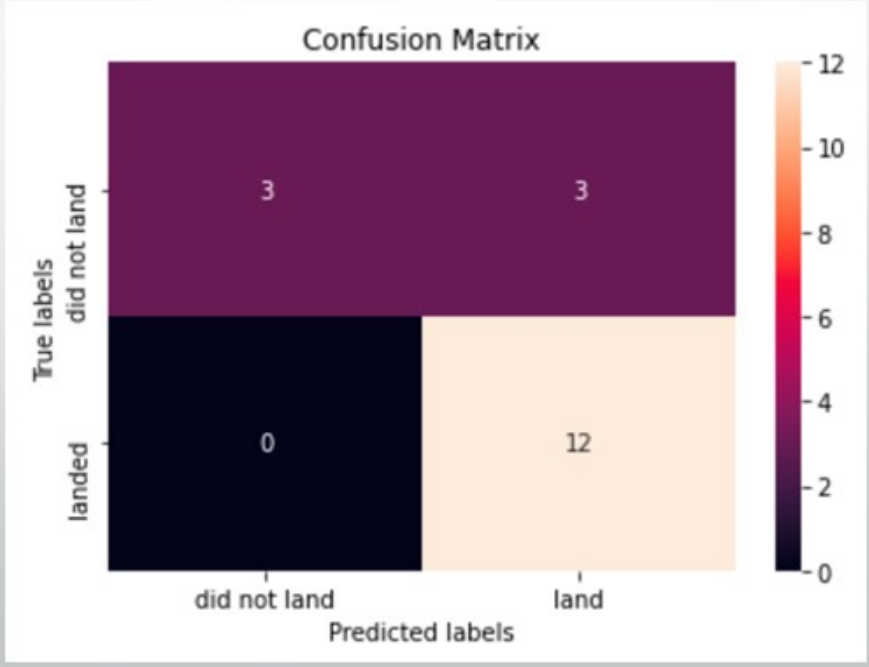


RESULTS : Dash

- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.

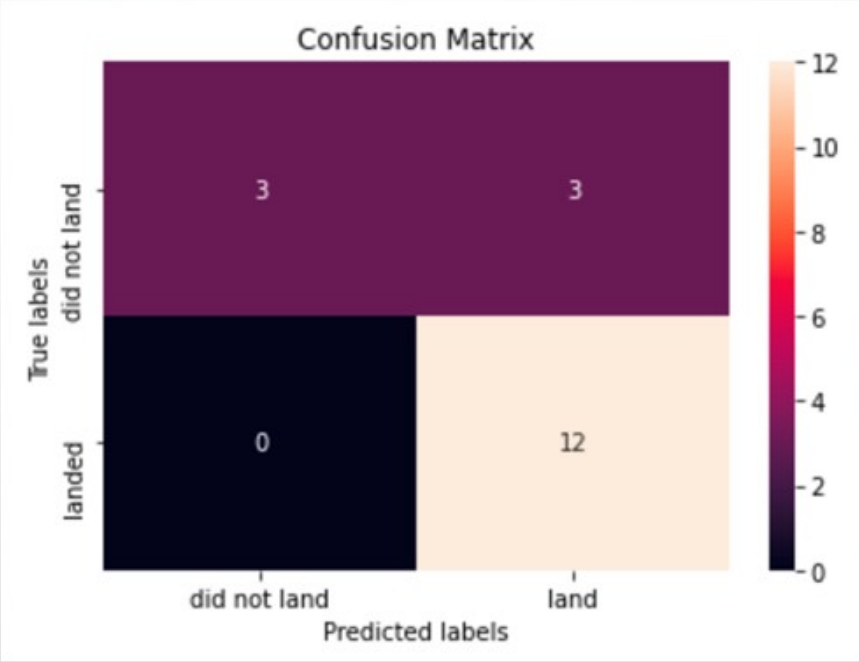
RESULTS : Predictive Analysis

- Logistic regression
  - GridSearchCV best score: 0.8464285714285713
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



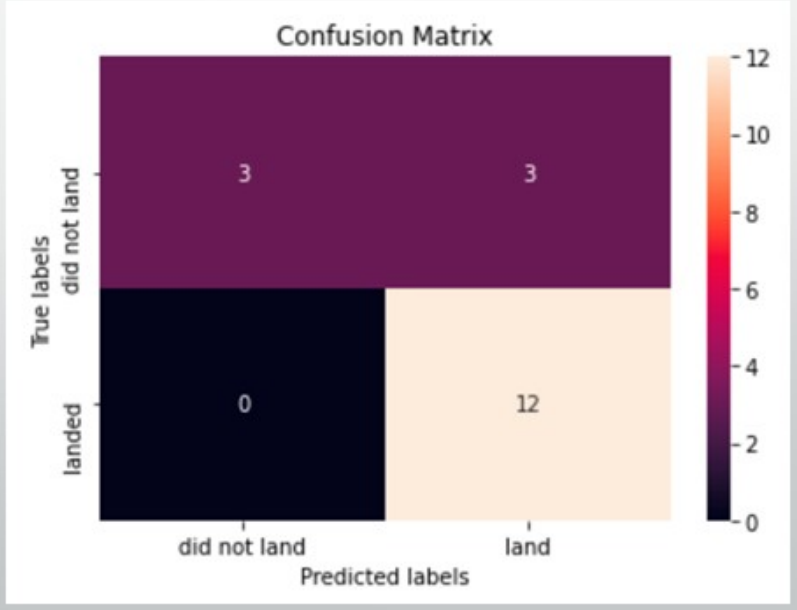
RESULTS : Predictive Analysis

- Support vector machine (SVM)
  - GridSearchCV best score: 0.8482142857142856
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



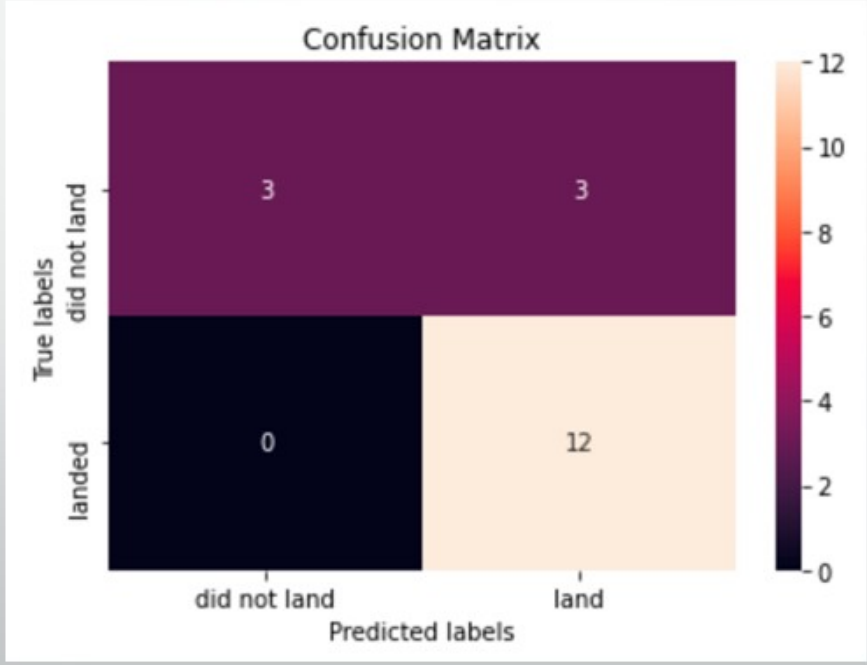
RESULTS : Predictive Analysis

- Decision tree
  - GridSearchCV best score: 0.8892857142857142
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS : Predictive Analysis

- K nearest neighbors (KNN)
  - GridSearchCV best score: 0.8482142857142858
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS : Predictive Analysis

- Putting the results of all 4 models side by side, we can see that they all share the same accuracy score and confusion matrix when tested on the test set.

- Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:

    1. Decision tree (GridSearchCV best score: 0.8892857142857142)
    2. K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
    3. Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
    4. Logistic regression (GridSearchCV best score: 0.8464285714285713)

DISCUSSION

Our visual analysis reveals distinct patterns in landing success rates:

1.Payload-Orbit Relationships:

1. Higher success rates observed for Polar, LEO, and ISS orbits with heavy payloads
2. GTO orbits show mixed outcomes regardless of payload mass

2.Feature Influence:

1. Multiple parameters appear correlated with mission success
2. Complex interactions make manual pattern recognition challenging

Machine Learning Approach:
To effectively model these relationships, we'll employ classification algorithms to:

•Learn historical success patterns

•Predict outcomes based on feature combinations

•Quantify each parameter's predictive importance

This data-driven approach will provide more reliable predictions than visual analysis alone.

CONCLUSION

This project develops a machine learning solution to predict whether SpaceX's Falcon 9 first stage will land successfully - a critical factor in determining launch costs. By analyzing various launch parameters (including payload mass and orbit type), we identify patterns that influence mission outcomes.

Our comparative analysis of multiple machine learning algorithms revealed that the Decision Tree model achieved superior performance in predicting landing success. This predictive capability enables more accurate cost estimation for Falcon 9 launches, providing valuable insights for the competitive space launch market.

The solution demonstrates how data science can extract meaningful business intelligence from complex aerospace operations.