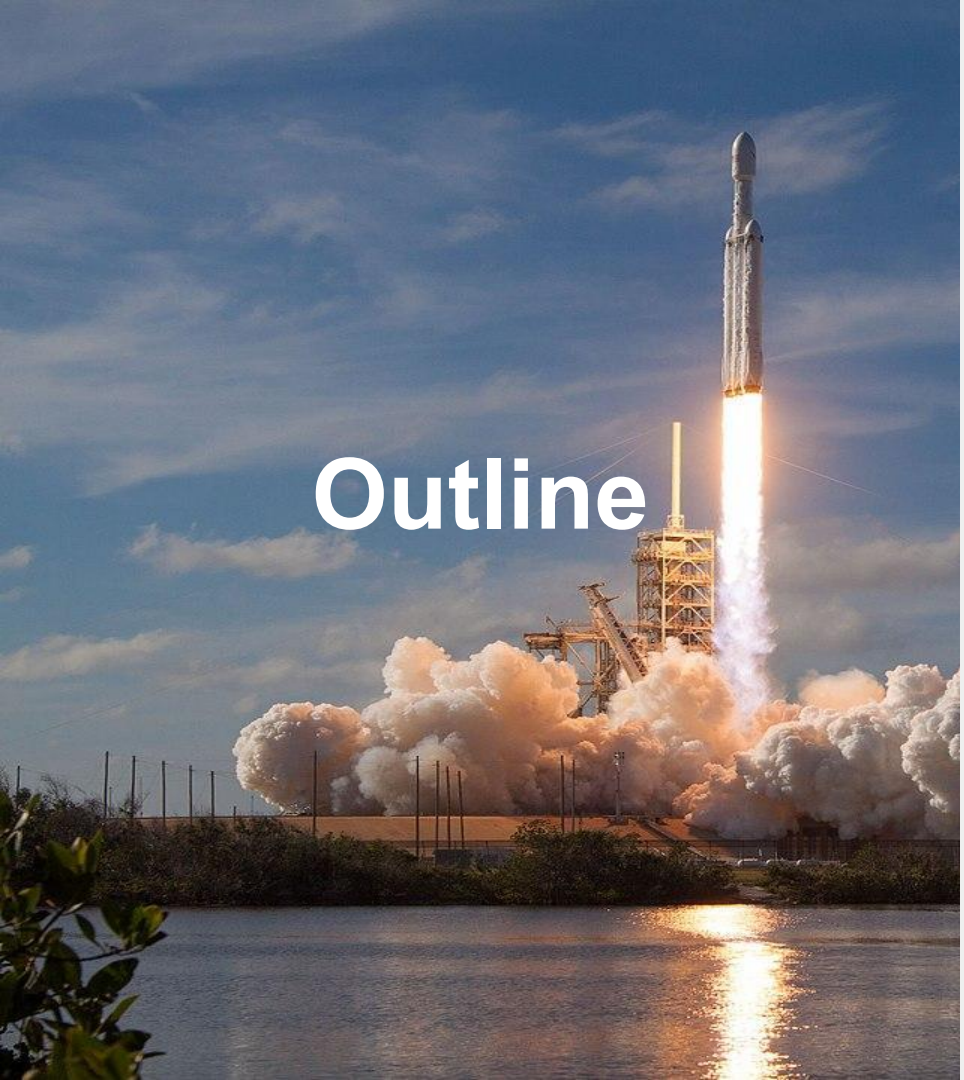
Anvith Thumma
July 12 2024

IBM Data Science Applied Capstone



Outline

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

Executive Summary

This capstone project aims to predict the successful landing of the SpaceX Falcon 9 first stage using various machine learning classification algorithms.

Project Highlights:

- **Data Collection, Wrangling, and Formatting**
- **Exploratory Data Analysis**
- **Interactive Data Visualization**
- **Machine Learning Prediction**

Key Findings:

- **Identified correlations between specific rocket launch features and landing outcomes (success or failure).**
- **Decision tree algorithm emerged as the most effective for predicting the successful landing of the Falcon 9 first stage.**

Introduction

In this capstone project, we aim to predict the successful landing of the Falcon 9 first stage. SpaceX offers Falcon 9 rocket launches significantly lower than other providers who charge upwards of $150 million. This cost reduction is largely due to SpaceX's ability to reuse the first stage. Predicting the first stage landing success can help determine the cost of a launch, providing valuable insights for alternate companies bidding against SpaceX.

Key Points:

- **Most unsuccessful landings are planned, often involving controlled landings in the ocean.**
- **The main question: Given features such as payload mass, orbit type, launch site, and more, can we predict if the Falcon 9 first stage will land successfully?**

Methodology

1. Data Collection, wrangling, and formatting:
   - SpaceX API
   - Web scraping
2. Exploratory data analysis (EDA):
   - Pandas and NumPy
   - SQL
3. Data Visualization:
   - Matplotlib and Seaborn
   - Folium
   - Dash
4. Machine learning Prediction:
   - Logistic regression
   - Support Vector machine (SVM)
   - Decision Tree
   - K-nearest neighbors (KNN)

Methodology (Data Collection, wrangling, and formatting)

SpaceX API:

- **API used: https://api.spacexdata.com/v4/launches/past**
- **The API provides data about many types of rocket launches, but we filtered for only Falcon 9 launches.**
- **Missing values are replaced by the mean of the specific column. We end up with 90 rows x 17 columns**

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False
5	8	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False
6	10	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False
7	11	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False
8	12	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False

Methodology (Data Collection, wrangling, and formatting)

Web Scraping:

- **Website used: https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922**
- **The data is later processed so that there are no missing entries and categorical features are encoded**
- **An extra column called Class is also added to the data frame. The column Class contains 0 if a given launch is failed and 1 if it is successful**
- **we end up with 90 rows x 83 columns**

	Flight No.	Launch site	Payload	Payload mass	Orbit
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO
1	2	CCAFS	Dragon	0	LEO
2	3	CCAFS	Dragon	525 kg	LEO
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO

Methodology (Exploratory data analysis (EDA))

Pandas and NumPy

- **Utilized for deriving basic information about the collected data:**
  - **Number of launches at each launch site**
  - **Number of occurrences of each orbit**
  - **Number and occurrences of each mission outcome**

SQL

- **Queried data to answer key questions:**
  - **Names of unique launch sites in the space mission**
  - **Total payload mass carried by boosters launched by NASA (CRS)**
  - **Average payload mass carried by booster version F9v1.1**

Methodology (Data Visualization)

Matplotlib and Seaborn

- **Utilized for visualizing data with scatterplots, bar charts, and line charts.**
- **Key relationships explored:**
  - **Flight number and launch site**
  - **Payload mass and launch site**
  - **Success rate and orbit type**

Folium

- **Used for creating interactive maps.**
- **Key features:**
  - **Mark all launch sites on a map**
  - **Indicate succeeded and failed launches at each site**
  - **Show distances between launch sites and nearby cities, railways, or highways**

Dash

- **Generates an interactive site with dropdown menus and range sliders.**
- **Visualizations include:**
  - **Total successful launches from each launch site (pie chart)**
  - **Correlation between payload mass and mission outcome (scatterplot) for each launch site**

Methodology (Machine learning Prediction)

Scikit-learn Library

- **Utilized for creating machine learning models.**

Prediction Phase Steps:

1. **Standardizing the Data**
2. **Splitting the Data**
    - **Training data**
    - **Test data**

1. **Creating Machine Learning Models**
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree
    - K-Nearest Neighbors (KNN)
2. **Model Training**
    - Fit the models on the training set
3. **Hyperparameter Tuning**
    - Find the best combination of hyperparameters for each model
4. **Model Evaluation**
    - Evaluate models based on accuracy scores and confusion matrix

Results

Sections:

1. **SQL (EDA with SQL)**
2. **Matplotlib and Seaborn (EDA with Visualization)**
3. **Folium**
4. **Dash**
5. **Predictive Analysis**

Key Insight:

- **In all following graphs:**
  - **Class 0: Failed launch outcome**
  - **Class 1: Successful launch outcome**

Results (SQL (EDA with SQL))

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

The names of the unique launch sites in the space mission

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

payloadmass
619967

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

5 records where launch sites begin with ‘CCA’

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

payloadmass
6138.287128712871

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

min(DATE)
2010-06-04

Results (SQL (EDA with SQL))

The total number of successful and failure mission outcomes

missionoutcomes
1
98
1
1

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

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

Results (SQL (EDA with SQL))

boosterversion
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

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

Results (SQL (EDA with SQL))

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

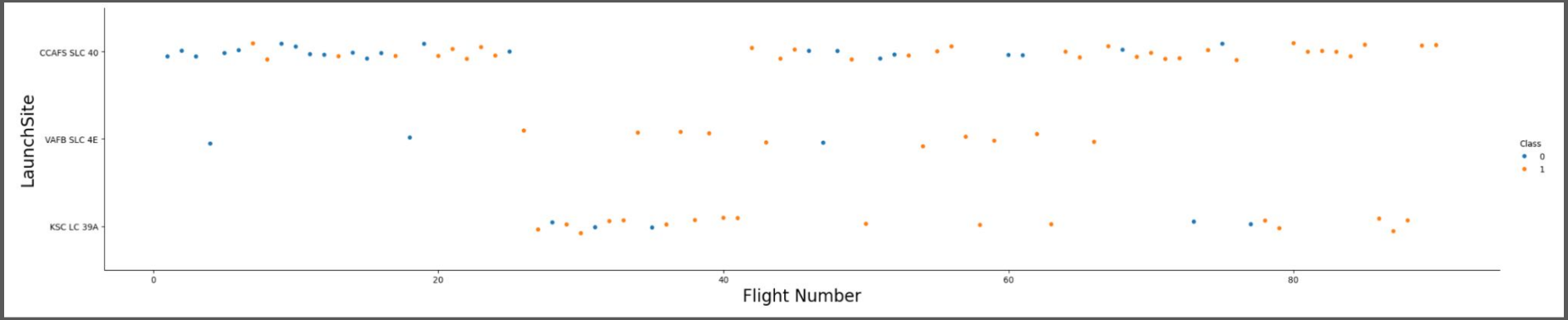
The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

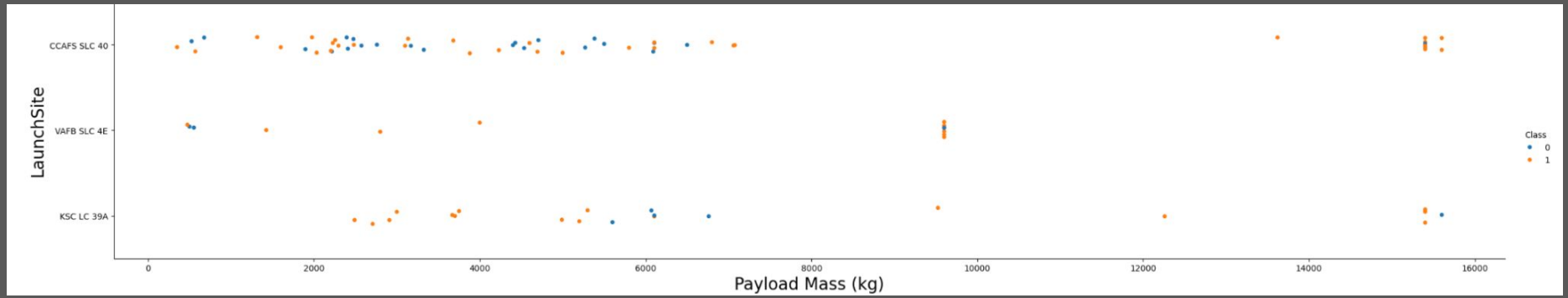
Results (Matplotlib and Seaborn)

The relationship between flight number and launch site



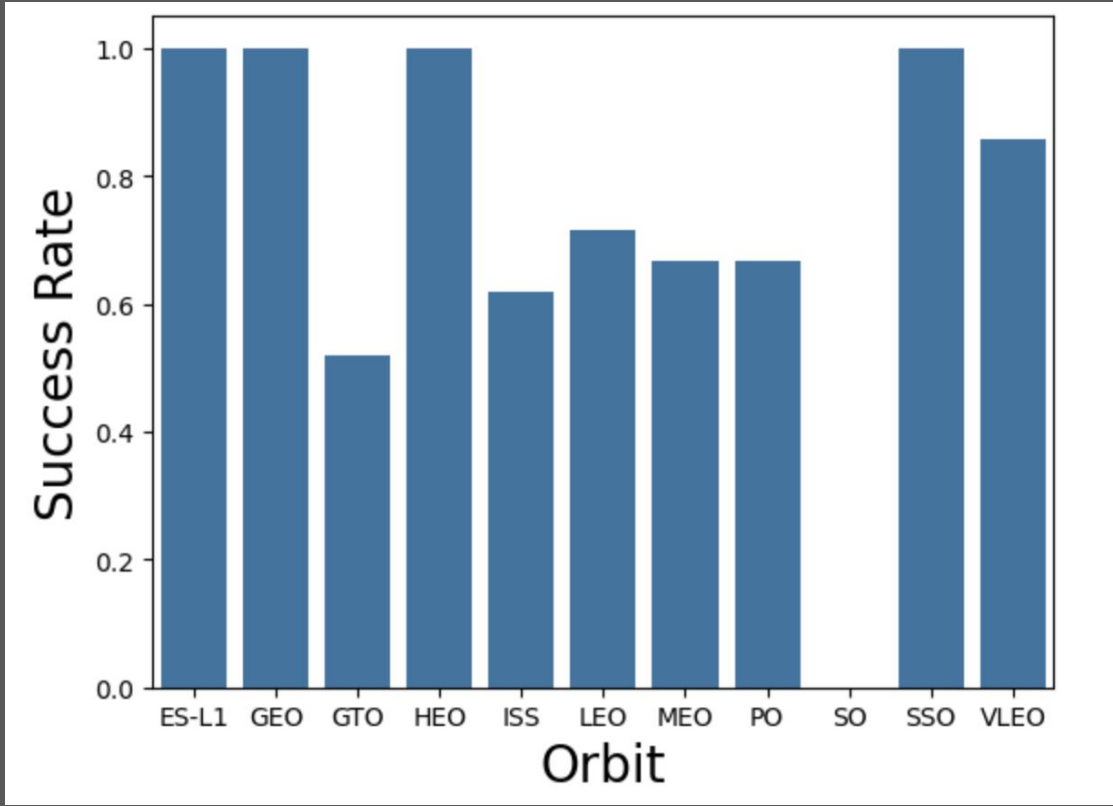
Results (Matplotlib and Seaborn)

The relationship between payload mass and launch site



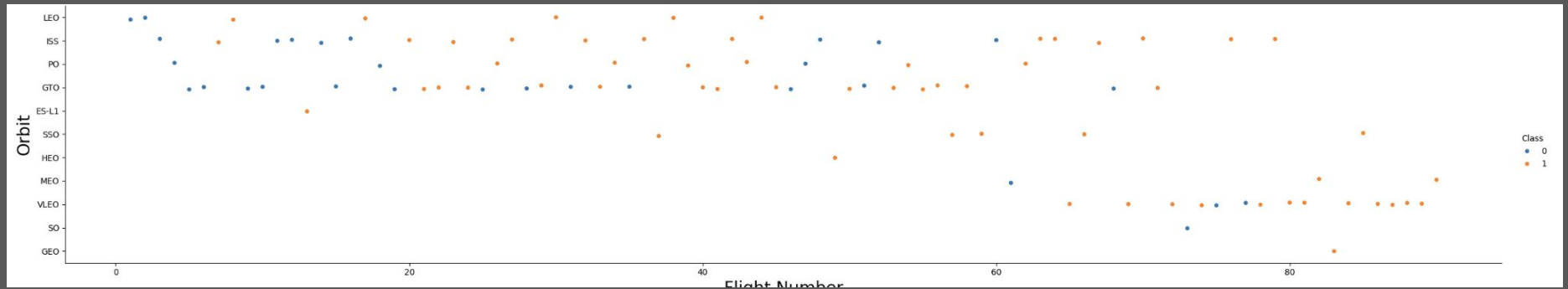
Results (Matplotlib and Seaborn)

The relationship between success rate and orbit type



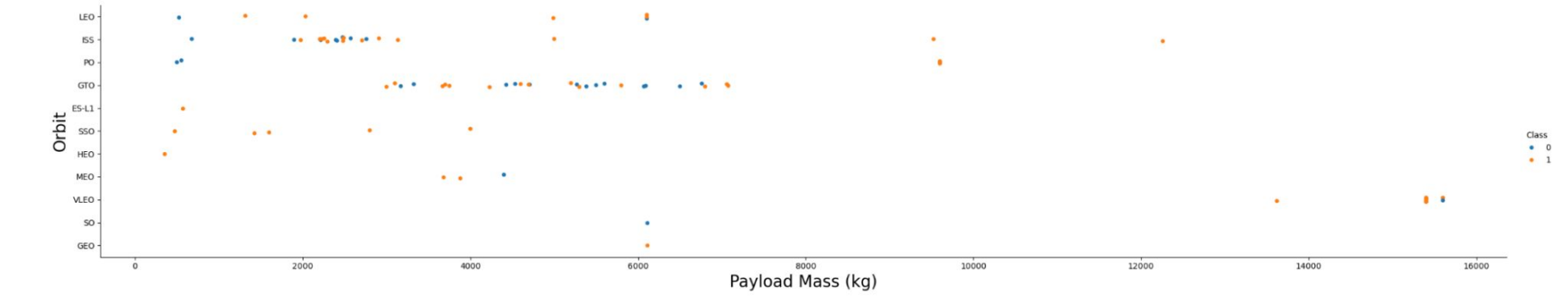
Results (Matplotlib and Seaborn)

The relationship between flight number and orbit type



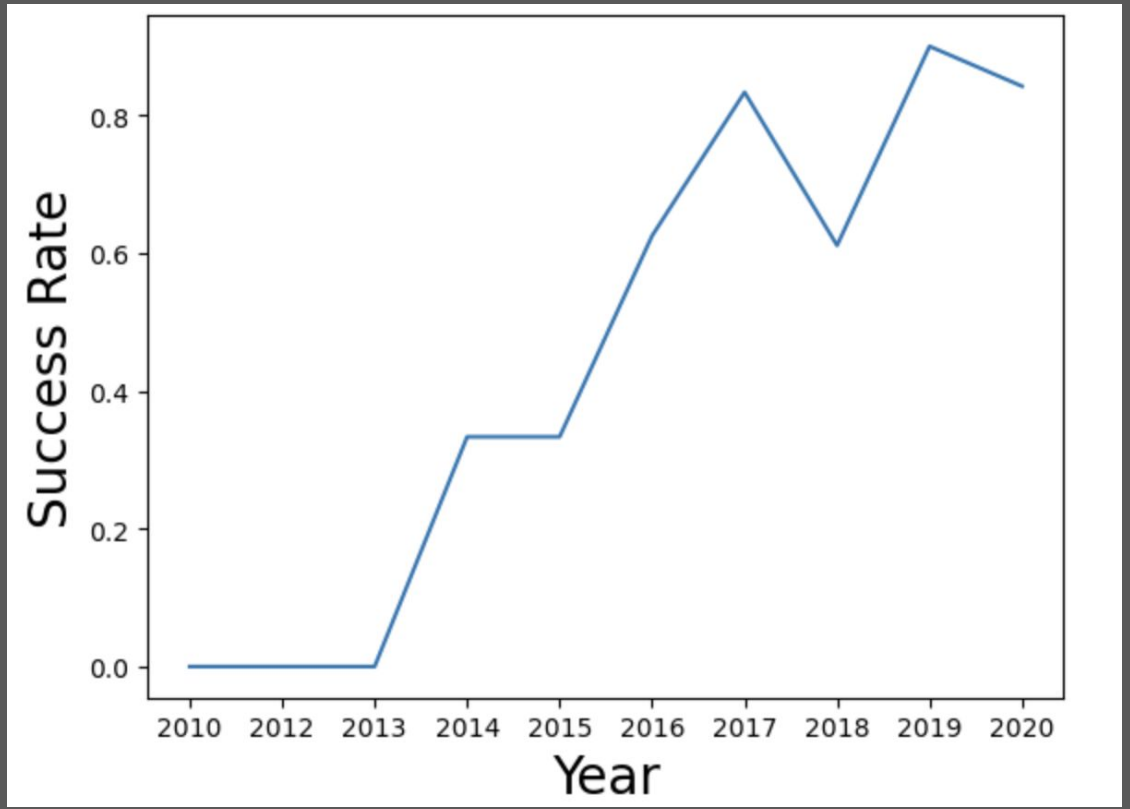
Results (Matplotlib and Seaborn)

The relationship between payload mass and orbit type



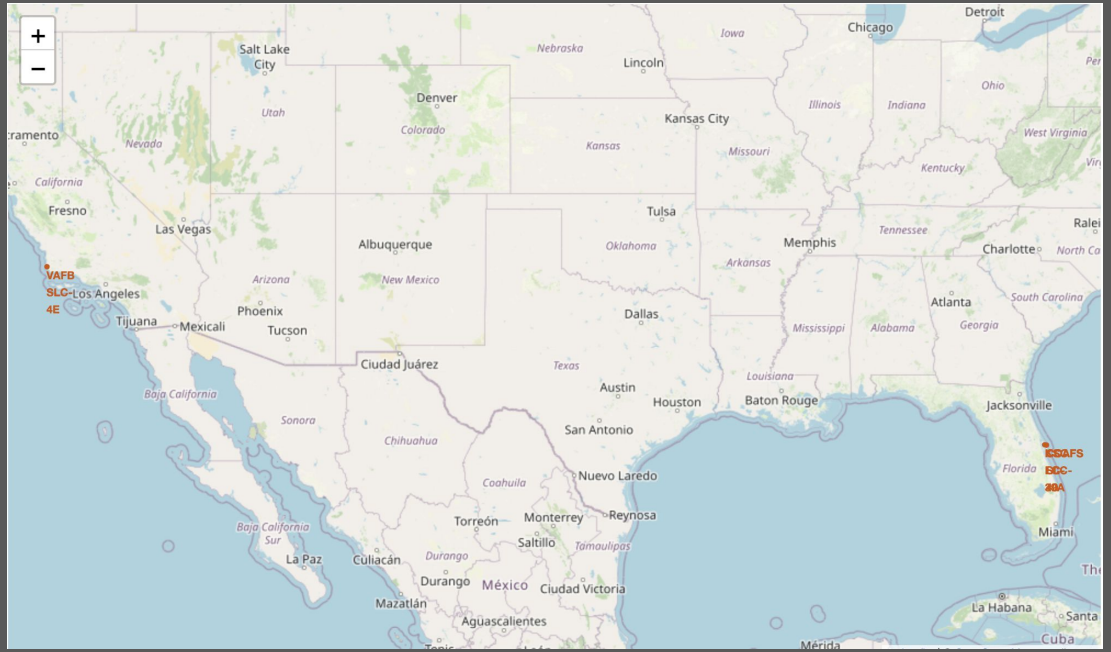
Results (Matplotlib and Seaborn)

The launch success yearly trend



Results (Folium)

All launch sites on map



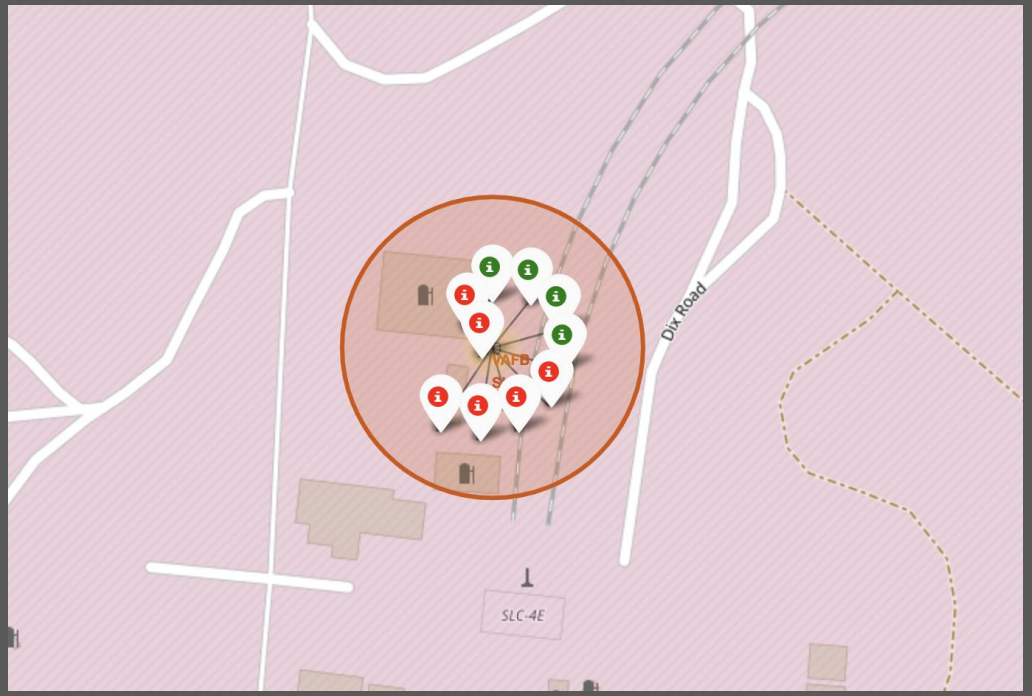
Results (Folium)

Key Features:

- **Succeeded and Failed Launches**
  - **Visualized for each site on the map**

Interactive Map:

- **Zoom in on a launch site to see:**
  - **Green Tags: Represent successful launches**
  - **Red Tags: Represent failed launches**



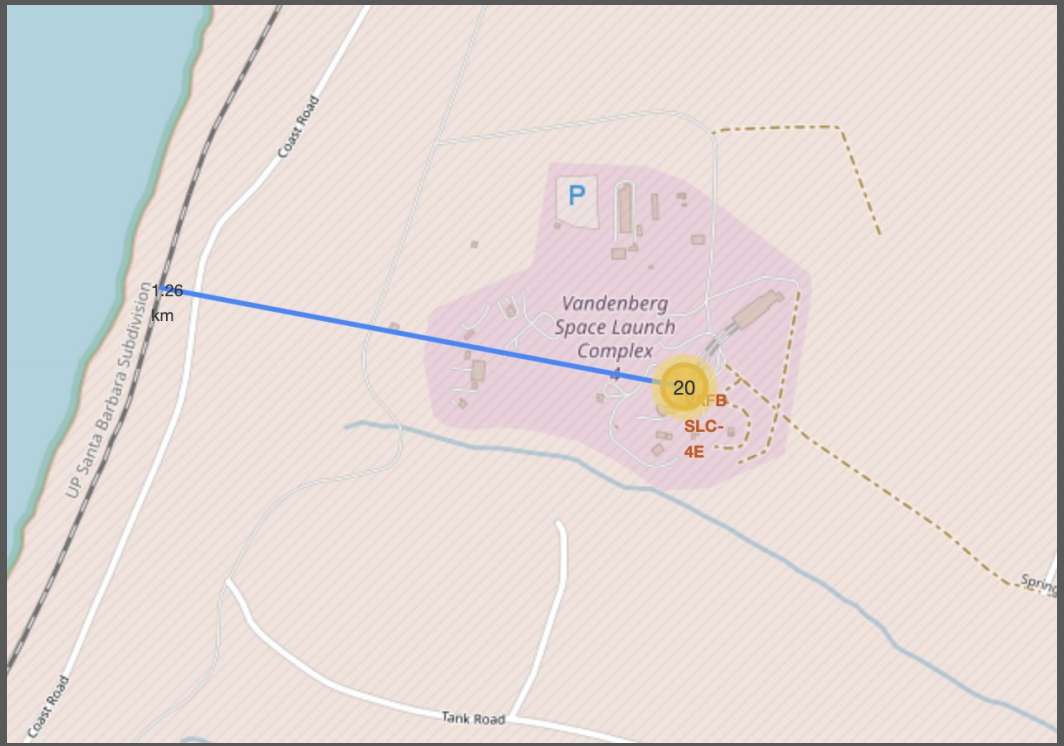
Results (Folium)

Key Measurements:

- **Distances between a launch site and its proximities:**
  - **Nearest city**
  - **Nearest railway**
  - **Nearest highway**

Example:

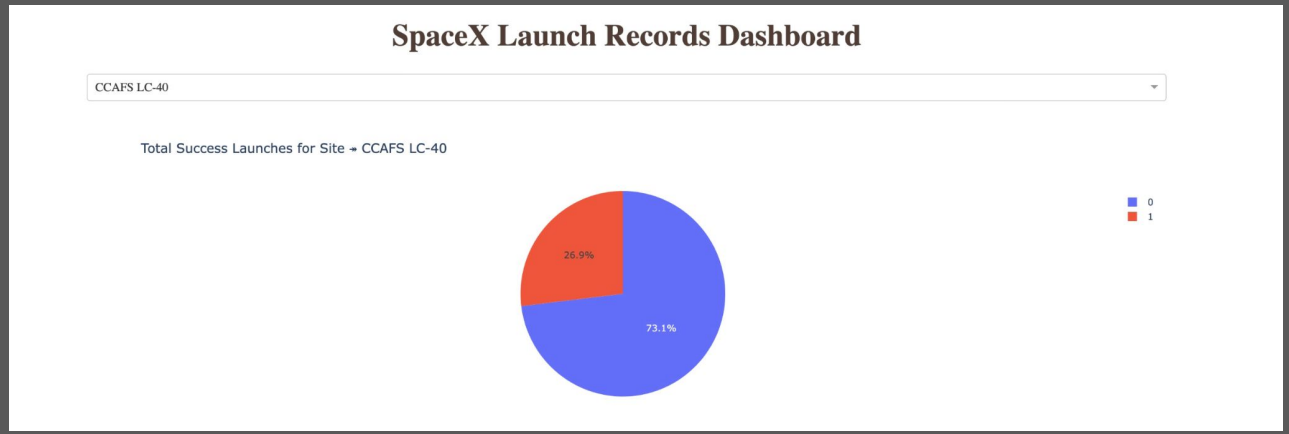
- **The image below shows the distance between the VAFB SLC-4E launch site and the nearest coastline:**



Results (Dash)

Pie Chart Visualization

- The image below displays the pie chart for launch site **CCAFS LC-40**.
- **Key Insights:**
  - **0**: Failed launches
  - **1**: Successful launches
  - **Findings**: 73.1% of launches at CCAFS LC-40 have resulted in failures.



Results (Dash)

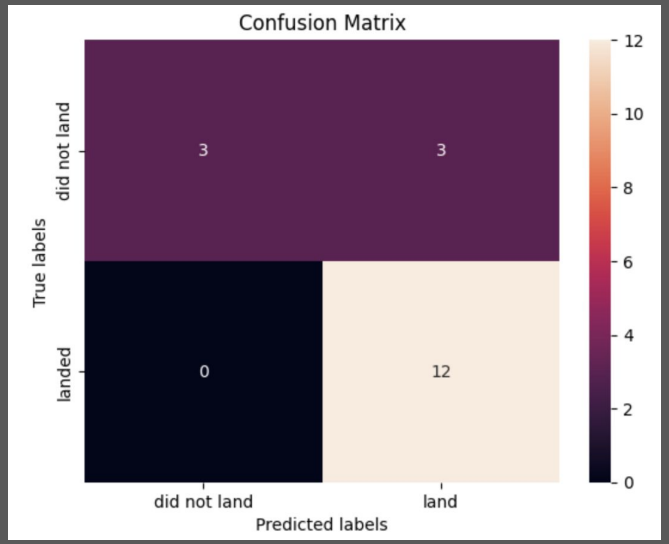
Scatterplot Visualization

- The image below shows the scatterplot for payload mass, with the range set from 2000 kg to 8000 kg.
- Key Insights:
  - Class 0: Represents failed launches
  - Class 1: Represents successful launches

Results (Predictive Analysis)

Logistic Regression Results

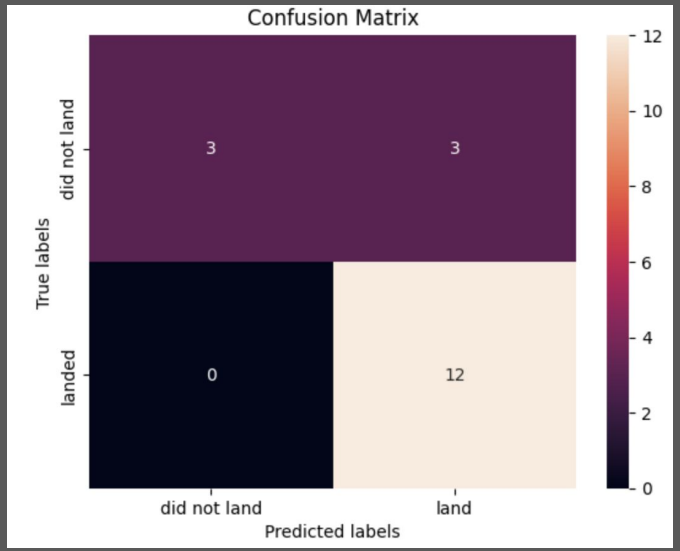
- **GridSearchCV Best Score: 0.8464285714285713**
- **Accuracy Score on Test Set: 0.8333333333333334**
- **Confusion Matrix:**



Results (Predictive Analysis)

Support Vector Machine (SVM) Results

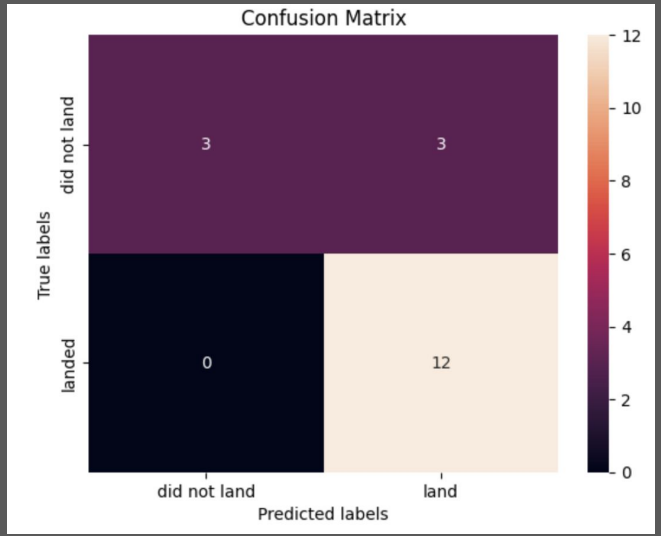
- **GridSearchCV Best Score: 0.8482142857142856**
- **Accuracy Score on Test Set: 0.8333333333333334**
- **Confusion Matrix:**



Results (Predictive Analysis)

Decision Tree Results

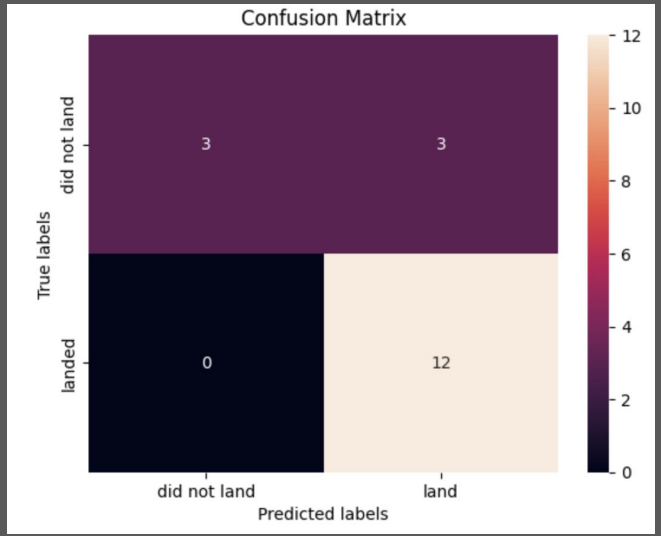
- **GridSearchCV Best Score: 0.875**
- **Accuracy Score on Test Set: 0.8333333333333334**
- **Confusion Matrix:**



Results (Predictive Analysis)

K-nearest neighbor (KNN) Results

- **GridSearchCV Best Score: 0.8482142857142858**
- **Accuracy Score on Test Set: 0.8333333333333334**
- **Confusion Matrix:**



Results (Predictive Analysis)

Model Comparison

Putting the results of all four models side by side, we observe that they share the same accuracy score and confusion matrix on the test set. Therefore, we rank them based on their GridSearchCV best scores:

Rank	Model	GridSearchCV Best Score
1	Decision Tree	0.875
2	K-nearest neighbor (KNN)	0.8482142857142858
3	Support Vector Machine (SVM)	0.8482142857142856
4	Logistic Regression	0.8464285714285713

Discussion

Feature Correlation and Impact on Mission Outcomes

From the data visualization section, we observe that certain features may correlate with mission outcomes in various ways:

- **Payload Impact: Heavy payloads tend to show higher successful landing rates for orbit types such as Polar, LEO, and ISS.**
- **GTO Complexity: For Geostationary Transfer Orbit (GTO), both successful and unsuccessful missions are present, making it difficult to distinguish trends.**

Key Insights:

- **Each feature appears to influence the final mission outcome in specific ways.**
- **The exact impact of these features can be challenging to interpret.**

Next Steps:

By employing machine learning algorithms, we can analyze historical data to identify patterns and predict the likelihood of mission success based on the given features.

Conclusion

Project Overview

In this project, we aim to predict whether the first stage of a Falcon 9 launch will successfully land, which helps determine the cost of the launch.

Key Points:

- **Feature Influence: Each feature of a Falcon 9 launch, such as payload mass or orbit type, may impact the mission outcome in specific ways.**
- **Machine Learning Approaches: Several machine learning algorithms are utilized to identify patterns in past Falcon 9 launch data, leading to predictive models for launch outcomes.**
- **Top Performer: Among the four machine learning algorithms employed, the predictive model generated by the decision tree algorithm demonstrated the best performance.**