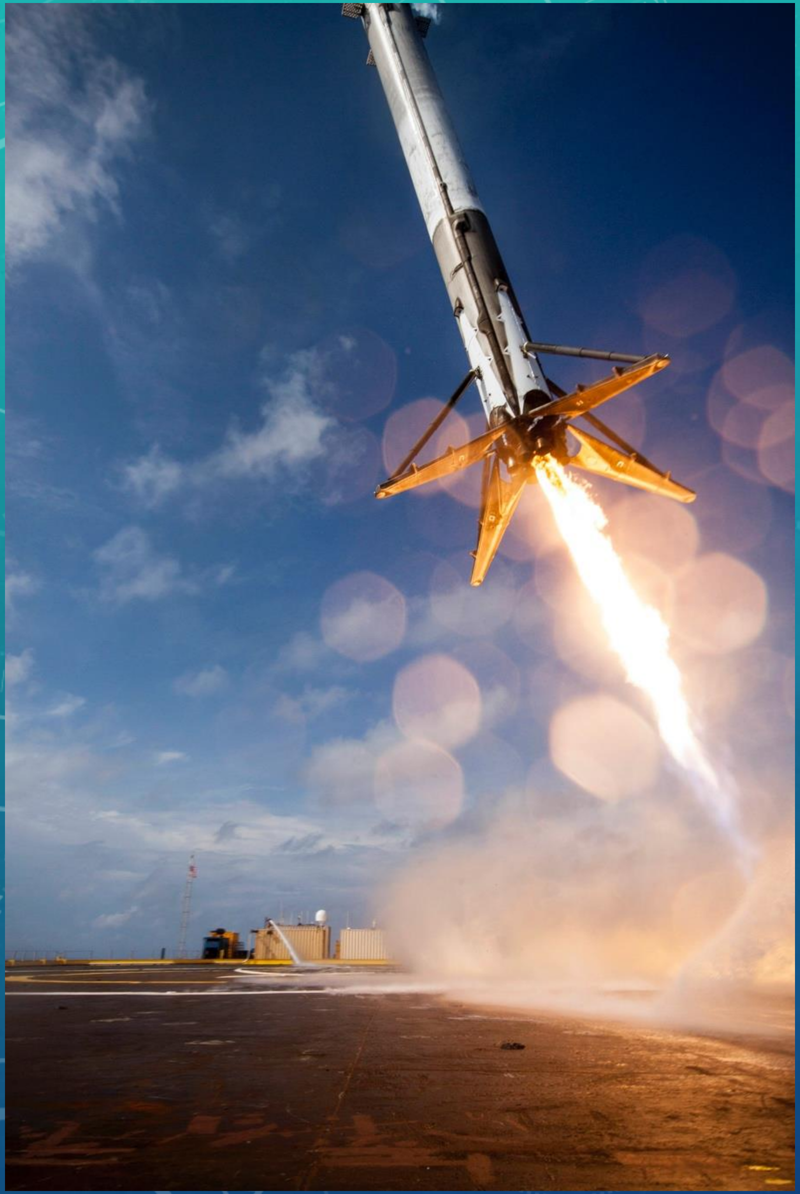


My first project as a data scientist: SPACE Y

Daniel Obed Ortega Vázquez

April 17, 2025

TABLE OF CONTENTS

EXECUTIVE SUMMARY

- **SpaceY** is a new commercial rocket launch provider aiming to compete with **SpaceX**.
- SpaceX offers launch services starting at **$62 million** for missions where sufficient fuel is reserved to enable recovery of the **first-stage rocket booster** for reuse.
- According to public statements, the cost to manufacture a Falcon 9 first-stage booster is estimated to exceed **$15 million**, excluding research and development costs or profit margins.
- This report presents models that, based on mission parameters such as payload mass and target orbit, can predict the likelihood of a successful first-stage landing with **83.3% accuracy**.
- These predictions provide **SpaceY** with a strategic advantage by allowing the company to estimate launch costs more precisely—enabling them to place **more competitive bids** against SpaceX.

INTRODUCTION

- **SpaceX** advertises **Falcon 9** rocket launches at a cost of **$62 million** when the **first stage** can be recovered and reused.
- The first stage alone is estimated to cost over **$15 million** to manufacture, excluding R&D recovery or profit margins.
- However, depending on mission parameters—such as **payload mass**, **target orbit**, or specific **customer requirements**—SpaceX may forgo recovering the first stage.
- As a result, this report seeks to accurately **predict the likelihood of a successful first-stage landing**, using it as a **proxy to estimate launch costs**.

METHODOLOGY

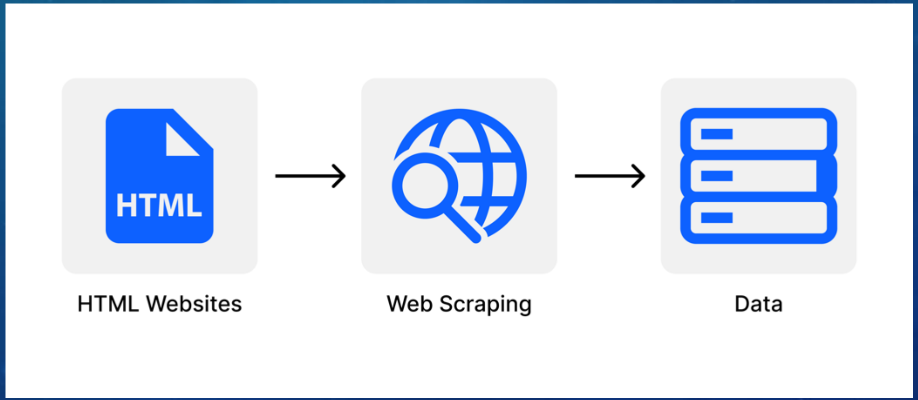
Data Collection

API

- Retrieved historical launch data from an open-source **SpaceX REST API**.
- Used **GET requests** to access and parse the SpaceX launch data.
- Filtered the dataset to include only **Falcon 9** launches.
- Handled missing values by replacing **classified mission payload masses** with the **mean payload mass** of known entries.

Web Scraping

- Collected additional launch data from the **Wikipedia page** titled *"List of Falcon 9 and Falcon Heavy Launches."*
- Accessed the page using its **Wikipedia URL**.
- Extracted all column names from the **HTML table headers**.
- Parsed the table and converted it into a structured **Pandas DataFrame** for analysis.

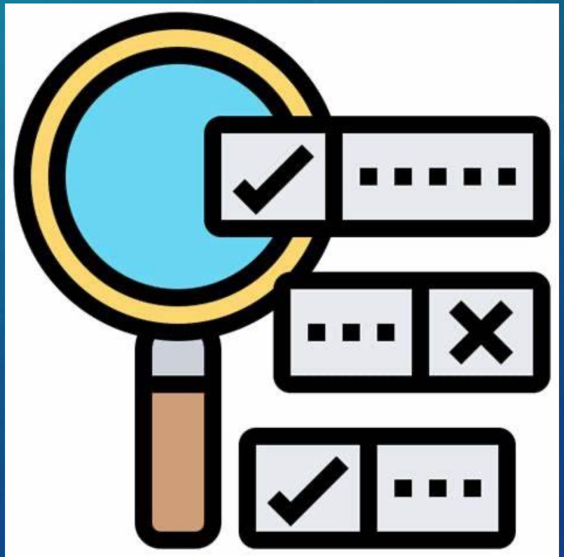


Data Wrangling

- Explored the dataset to identify an appropriate **target label** for training supervised learning models.
- Performed exploratory analysis including:
  - Count of launches per **launch site**
  - Frequency and distribution of each **orbit type**
  - Analysis of **mission outcomes** by orbit category
- Created a binary **training label** named 'Class', derived from the 'Outcome' column, to indicate the **success or failure of first-stage booster landings**:

Landing Outcome Labeling:

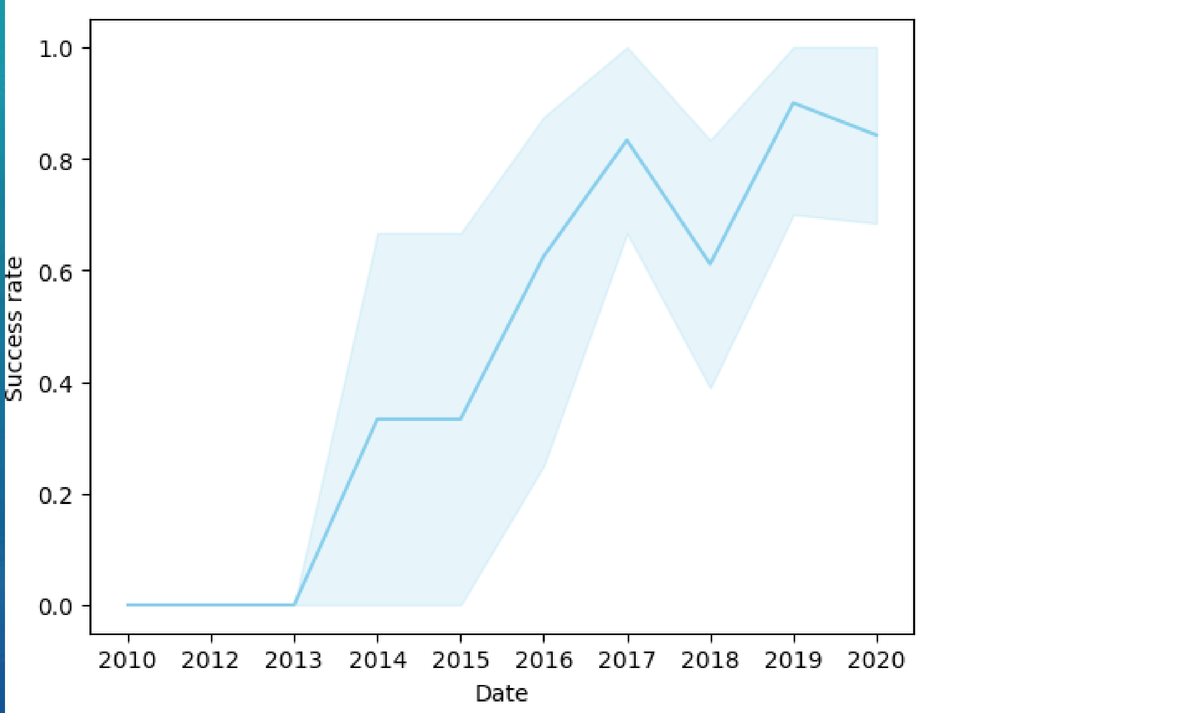
- **Class = 0** → First-stage booster **did not land successfully**:
  - 'None None': Landing **not attempted**
  - 'None ASDS': **Attempt not possible** due to launch failure
  - 'False ASDS': **Drone ship landing failed**
  - 'False Ocean': **Ocean landing failed**
  - 'False RTLS': **Ground pad landing failed**
- **Class = 1** → First-stage booster **landed successfully**:
  - 'True ASDS': **Drone ship landing succeeded**
  - 'True RTLS': **Ground pad landing succeeded**
  - 'True Ocean': **Ocean landing succeeded**



Exploratory Data Analysis (EDA)

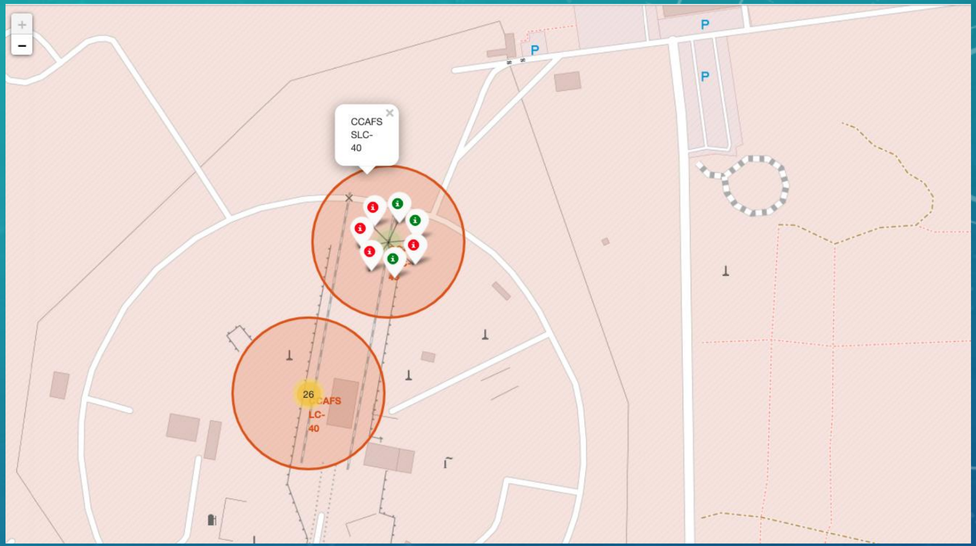
- **EDA with SQL**
  - Loaded the dataset into an **IBM Db2** instance.
  - Executed **SQL queries** to explore and retrieve insights on:
    - Launch sites
    - Payload masses
    - Booster versions
    - Mission outcomes
    - Booster landings
- **EDA with Visualization**
  - Imported the dataset into a **Pandas DataFrame** for visual analysis.
  - Used **Matplotlib** and **Seaborn** to create visualizations for deeper insight:
    - Flight Number vs. Payload Mass
    - Flight Number vs. Launch Site
    - Payload Mass vs. Launch Site
    - Orbit Type vs. Landing Success Rate
    - Flight Number vs. Orbit Type
    - Payload Mass vs. Orbit Type
    - Launch Year vs. Success Rate
- *(†) Plots used to identify trends and relationships between mission parameters and launch outcomes.*

```python
# Plot a line chart with x axis to be the extracted year and y axis to be the success rate
sns.lineplot(data=df, x='Date', y='Class', estimator='mean', color='skyblue')
plt.ylabel('Success rate')
plt.xlabel('Date')
plt.show()
```

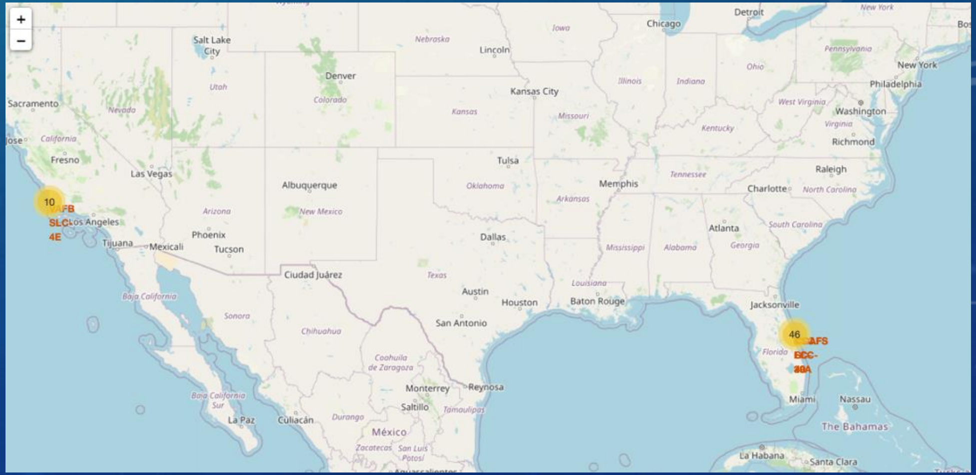


Data Visualization

- **Launch Sites Location Analysis**
  - Utilized the **Folium** Python library for **interactive geographic mapping**.
  - Mapped the **locations of all SpaceX launch sites**.
  - Plotted **individual launch markers** indicating **success or failure** at each site.
  - Calculated distances from each launch site to nearby infrastructure and landmarks:
    - **Railways**
    - **Highways**
    - **Coastlines**
    - **Cities**

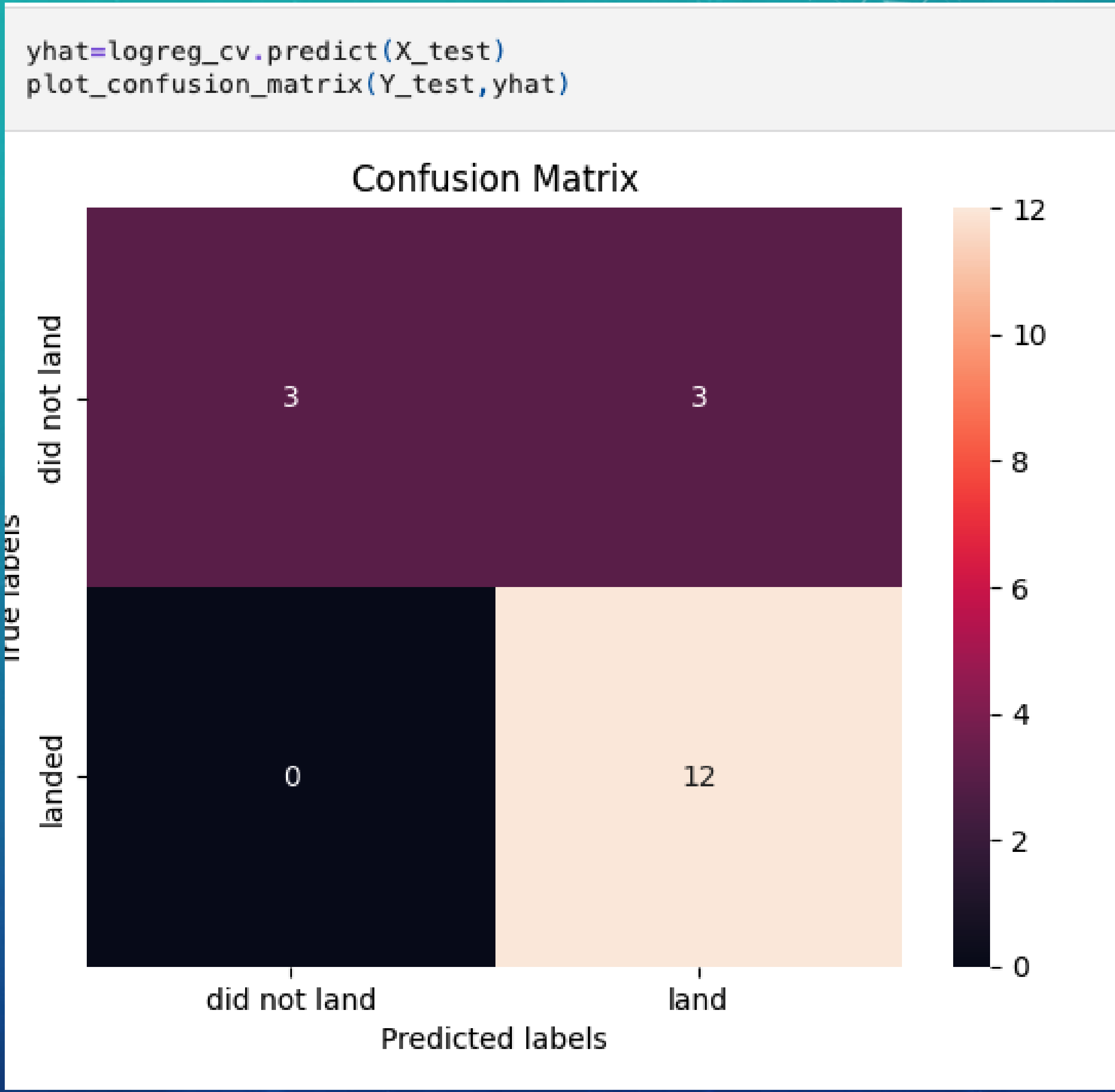


- **Launch Records Dashboard**
  - Built an **interactive dashboard** using **Plotly Dash** to allow real-time data exploration for stakeholders.
  - Key features include:
    - **Pie chart** displaying launch **success rates**, color-coded by **launch site**
    - **Scatter plot** of **Payload Mass vs. Landing Outcome**, color-coded by **booster version**
      - Includes a **range slider** to filter by payload mass
      - Features a **dropdown menu** to select between **all sites** or **individual launch sites**
  - Deployed the dashboard as a **static web application** using **Heroku**:
    IBM Applied Data Science Capstone Dashboard



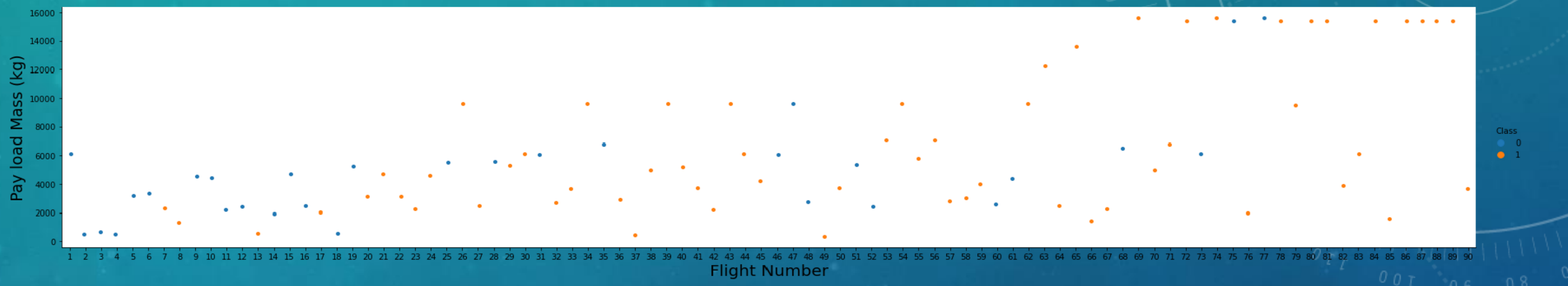
Predictive Analysis (Modeling)

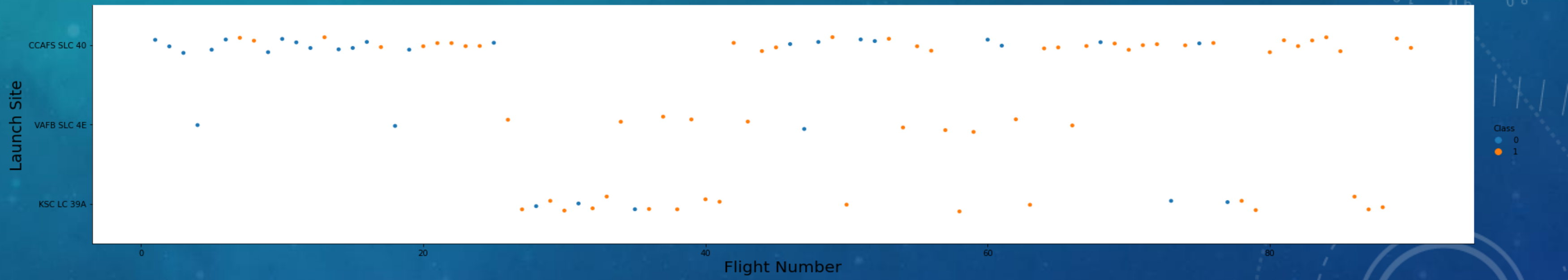
- Imported essential libraries and defined a function to generate a **confusion matrix**:
    - pandas
    - numpy
    - matplotlib
    - seaborn
    - scikit-learn (sklearn)
- Loaded the cleaned **DataFrame** created during the data collection phase.
- Added the **'Class'** column, previously created during data wrangling, as the **target label**.
- **Standardized the features** to ensure uniformity in model training.
- **Split** the dataset into **training** and **test sets**.
- **Model Training**
- Trained and evaluated the following classification models:
    - **Logistic Regression**
    - **Support Vector Machine (SVM)**
    - **Decision Tree Classifier**
    - **K-Nearest Neighbors (KNN) Classifier**
- **Model Optimization**
- Performed **cross-validated grid search** to find the **optimal hyperparameters** for each model.
- Utilized **Scikit-learn's GridSearchCV** for automated hyperparameter tuning.
- **Model Evaluation**
- Assessed **model accuracy** on the test set.
- Selected the **best-performing model** based on evaluation metrics, including accuracy and confusion matrix results.

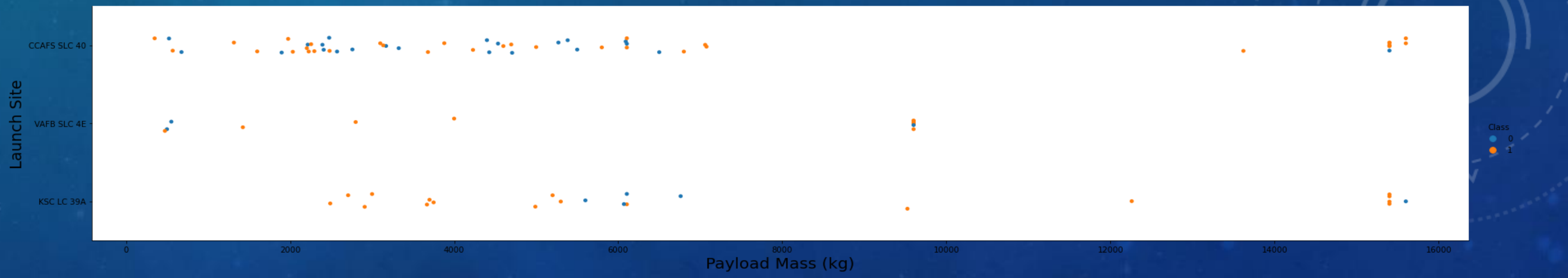


EDA WITH VISUALIZATION RESULTS

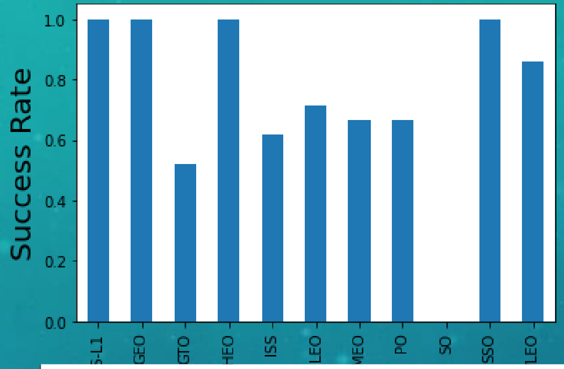
- **Flight Number vs. Payload Mass**
- **First-stage landing success** shows a **positive correlation** with the **number of launch attempts** (FlightNumber), suggesting improvement over time.
- Conversely, there is a **negative correlation** with **payload mass**, indicating that **heavier payloads** are associated with a **lower likelihood of successful landings**.
- **Flight Number vs. Launch Site**
- Early **first-stage landing failures** were primarily associated with **CCAFS SLC-40**, indicating this site was heavily used during the initial testing and development phase.
- **Payload Mass vs. Launch Site**
- **CCAFS SLC-40** and **KSC LC-39A** appear to be the preferred sites for **heavier payload missions**, possibly due to their infrastructure and proximity to the ocean, which facilitates risk mitigation.

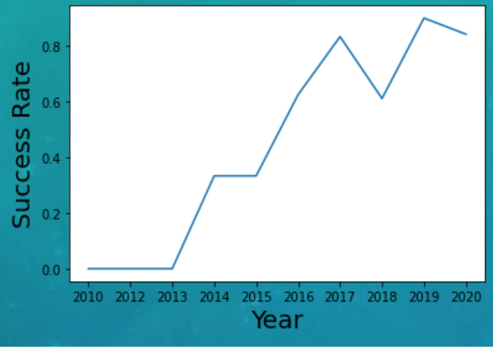




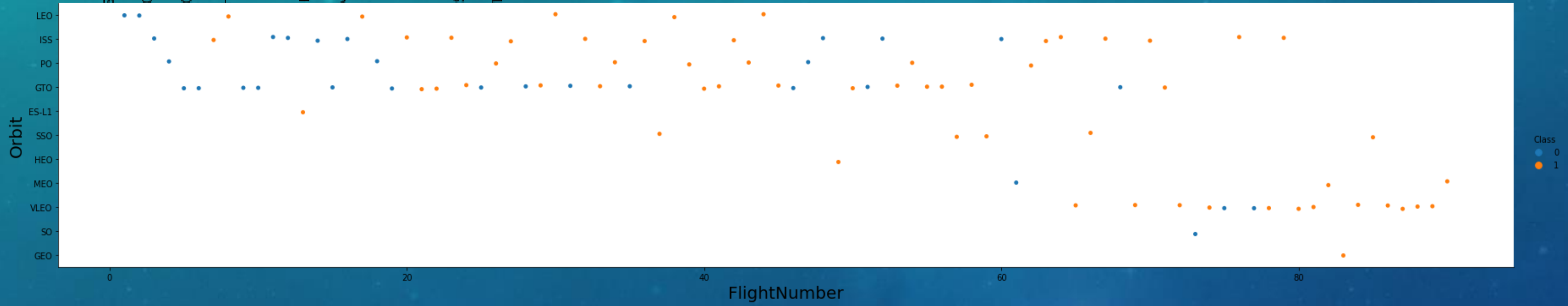


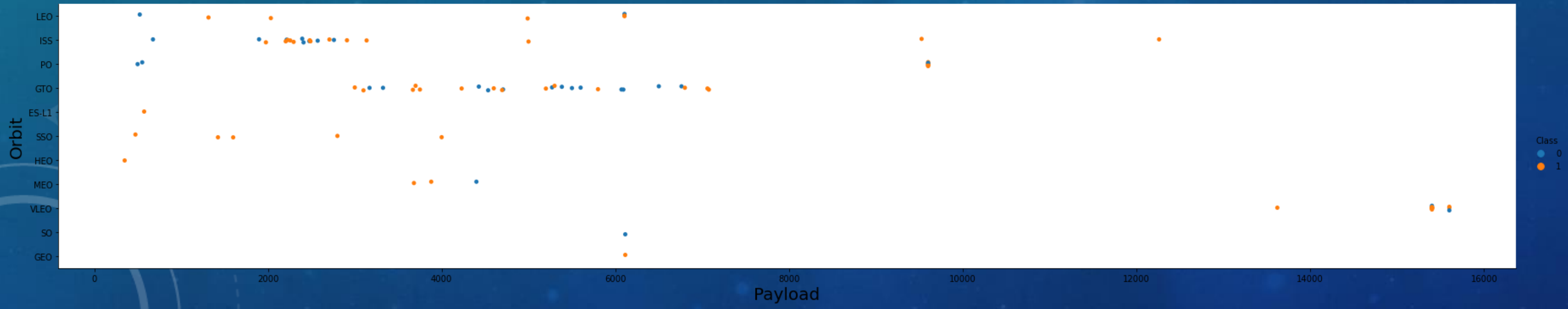
- ## Year vs. Success Rate
  - The **success rate of first-stage landings** has shown a **positive trend year over year** since **2013**, reflecting continuous improvements in technology, mission planning, and landing techniques.





- ## Flight Number vs. Orbit Type
  - Across all **orbit types**, there is a **positive correlation** between **flight number** and **first-stage recovery**, suggesting that **SpaceX has become increasingly successful at landing boosters** as launch experience accumulates.





- ## Payload Mass vs. Orbit Type
  - For **Geostationary Transfer Orbit (GTO)** missions, **heavier payloads** tend to **reduce the likelihood of successful first-stage recovery**, likely due to fuel constraints.
  - In contrast, for **International Space Station (ISS)** missions, **heavier payloads** are **positively associated** with successful landings—possibly because these missions typically target **lower orbits**, allowing for more fuel to be reserved for landing.

EDA WITH SQL RESULTS

- The team at SpaceY set out to answer several targeted questions using SQL:
- **What launch sites has SpaceX used?**
  - CCAFS LC 40
  - CCAFS SLC 40
  - KSC LC 39A
  - VAFB SLC 4E
- **Do launch site and date records for sites beginning with 'CCA' overlap?**
  - The **last launch** from *CCAFS LC 40* occurred on **2016-08-14**.
  - The **first launch** from *CCAFS SLC 40* occurred on **2017-12-15**.
  - According to Wikipedia, *Cape Canaveral Space Launch Complex 40* was renamed in **2017**.
- **What is the total payload mass carried by boosters launched under NASA's CRS (Commercial Resupply Services) program?**
  - **45,596 kg** total.
- **What is the average payload mass carried by the F9 v1.1 booster version?**
  - **340 kg** on average.
- **When was the first successful landing on a ground pad achieved?**
  - On **2015-12-22**, more than five years after the first Falcon 9 launch on **2010-06-04**.

1. Boosters with Successful Drone Ship Landings

Criteria: Payload mass between 4,000 kg and 6,000 kg

Boosters:

•F9 FT B1021.1
•F9 FT B1023.1
•F9 FT B1029.2
•F9 FT B1038.1
•F9 B4 B1042.1
•F9 B4 B1045.1
•F9 B5 B1046.1

2. Landing Outcomes Count (from 2010-06-04 to 2017-03-20)

Ranked in descending order:

1.10 — *No attempt*
2.5 — *Failure (drone ship)*
3.5 — *Success (drone ship)*
4.3 — *Controlled (ocean)*
5.3 — *Success (ground pad)*
6.2 — *Failure (parachute)*
7.2 — *Uncontrolled (ocean)*
8.1 — *Precluded (drone ship)*

3. Booster Versions with Maximum Payload Mass

Boosters:

•F9 B5 B1048.4
•F9 B5 B1048.5
•F9 B5 B1049.4
•F9 B5 B1049.5
•F9 B5 B1049.7
•F9 B5 B1051.3
•F9 B5 B1051.4
•F9 B5 B1051.6
•F9 B5 B1056.4
•F9 B5 B1058.3
•F9 B5 B1060.2
•F9 B5 B1060.3

4. Failed Drone Ship Landings in 2015

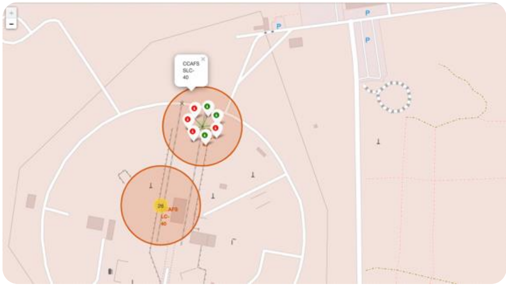
Landing Outcome – Booster Version – Launch Site:

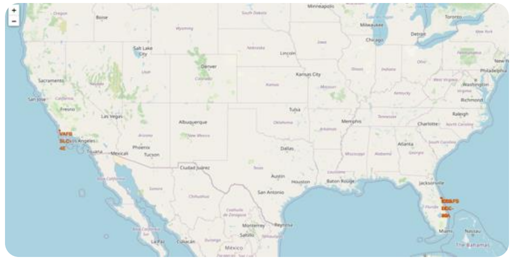
•Failure (drone ship) — F9 v1.1 B1012 — CCAFS LC 40
•Failure (drone ship) — F9 v1.1 B1015 — CCAFS LC 40

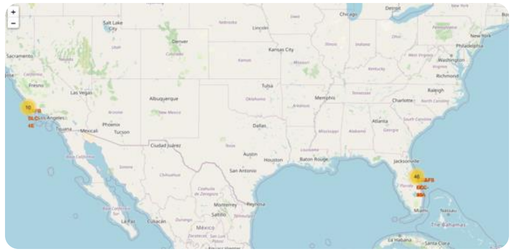
5. Mission Outcomes (Total Count)

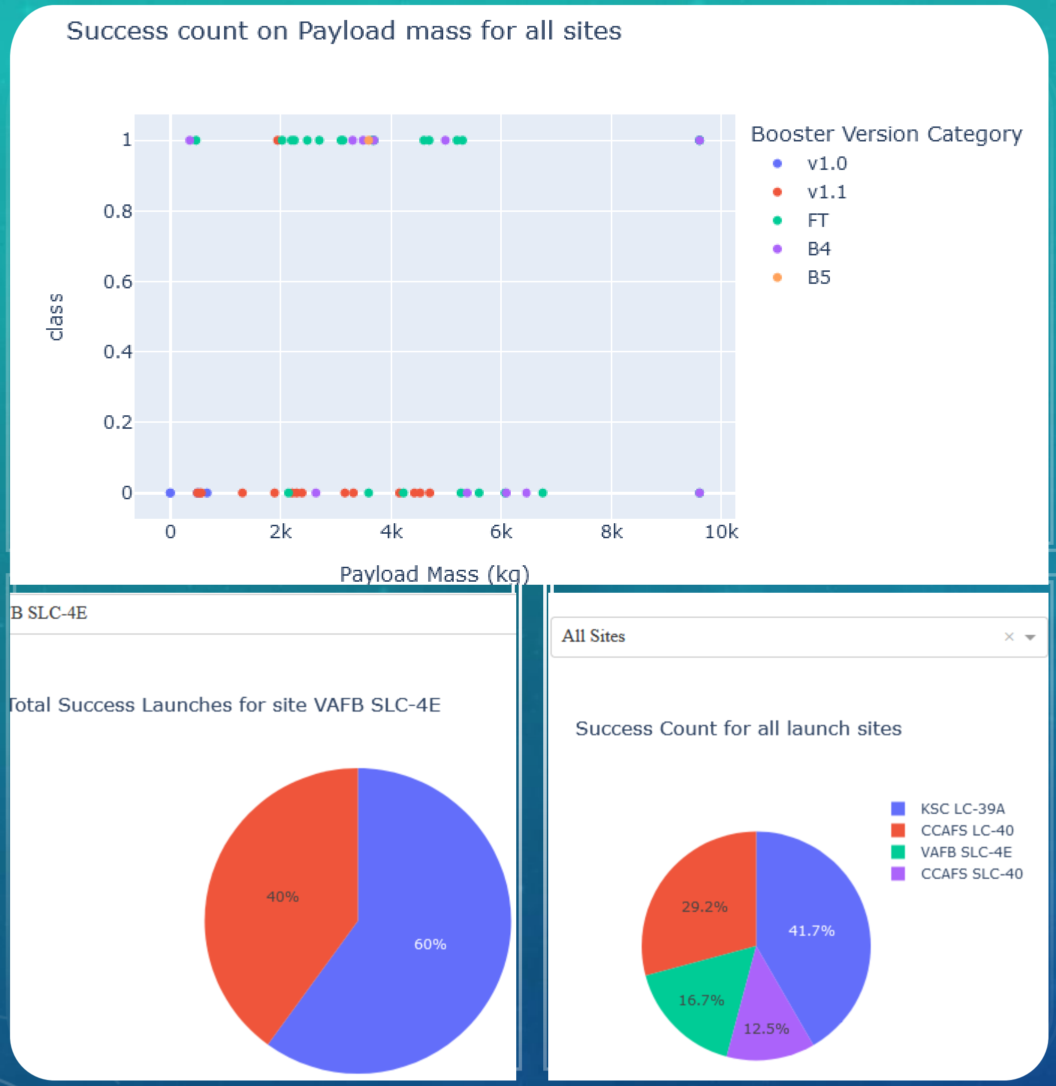
•99 — *Success*
•1 — *Failure (in flight)*
•1 — *Success (payload status unclear)*

INTERACTIVE MAP WITH FOLIUM RESULTS









PLOTLY DASH DASHBOARD RESULTS

- **Explore the Dashboard**
- Stakeholders can explore and interact with the data in real time via the interactive dashboard:
  IBM Applied Data Science Capstone Dashboard
- The dashboard enables users to visualize trends, filter data, and gain insights dynamically.
- **Key Observations from the Dashboard:**
- **VAFB SLC 4E** recorded the **heaviest successful booster landing**.
- **KSC LC 39A** has the **highest booster landing success rate** among all launch sites.
- **Payloads under 5,300 kg** are associated with the **highest landing success rates**.
- In contrast, **payloads over 5,300 kg** tend to have the **lowest booster landing success rates**.

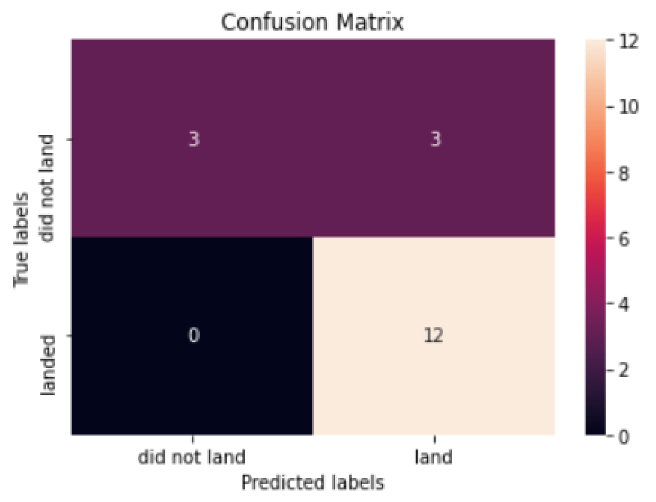
PREDICTIVE ANALYSIS (CLASSIFICATION) RESULTS

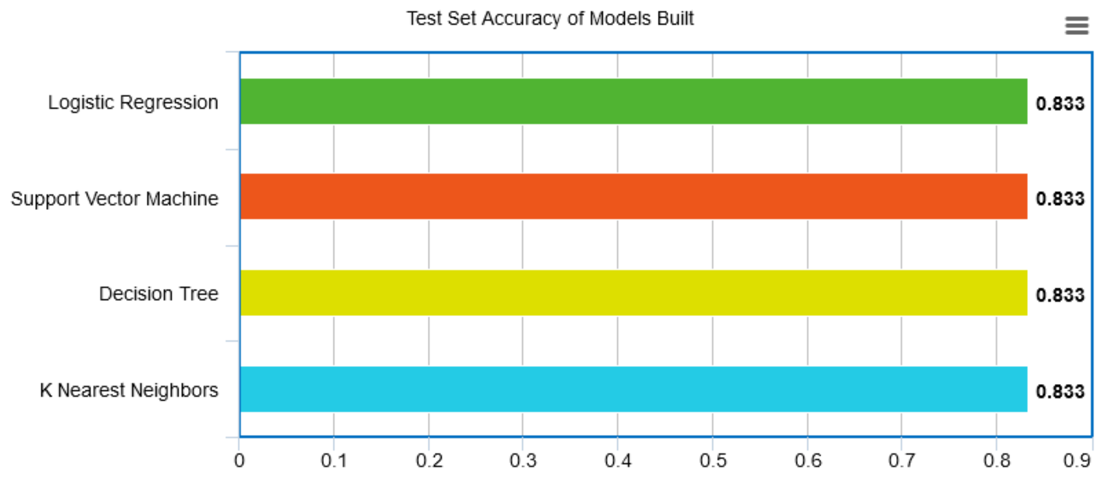
Model Performance Summary

- All **four models** achieved the **same accuracy score**: **83.33%**.
- The **confusion matrices** for the top-performing models were **identical**, indicating a **four-way tie** in performance.

Key Issue: False Positives

- A major issue across the models is the **false positive rate**.
- Specifically, the models **incorrectly predicted a successful landing** for the 1st stage booster in **3 out of 18 test samples**.
- This suggests a tendency to **overpredict successful landings**, which could lead to **misleading performance expectations** in real-world applications.





CONCLUSIONS AND STRATEGIC RECOMMENDATIONS

Key Findings

- Using the models developed in this report, **SpaceY can predict with 83.3% accuracy** whether SpaceX will **successfully land the 1st stage booster**.
- According to public statements, **SpaceX reports that each 1st stage booster costs over $15 million** to manufacture.
- This predictive capability gives SpaceY a competitive edge, allowing it to **make more informed and strategic bids** when competing with SpaceX.
- If SpaceX **fails to recover the 1st stage**, the total launch cost could rise from the **list price of $62 million** to approximately **$77 million**, factoring in the **$15M+ loss**.

Opportunities for Improvement and Future Work

1. **Finalize and Retrain Best Model**
    1. **Freeze** the best-performing model and hyperparameters.
    2. **Retrain using the full dataset** (training + test) to improve model generalization.

    **Note*: This approach enhances predictive power but eliminates the ability to re-evaluate test accuracy.
2. **Incorporate Additional Launch Data**
    1. Continuously update the model as **new SpaceX launch data** becomes available to improve accuracy and relevance.
3. **Subdivide the Prediction Task**
    1. Split the current model into two sequential predictions:
        1. **Will SpaceX attempt** to land the 1st stage booster?
        2. **Will SpaceX succeed** in that attempt?
    2. This granularity can help SpaceY evaluate both **risk and cost implications** more accurately.
4. **Develop a Booster Reuse Model**
    1. Create a related model to **predict whether SpaceX will use a previously flown 1st stage**.
    2. This would help anticipate when **SpaceX bids might include a discount**, giving SpaceY a **clearer pricing forecast**.