IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Han Xu
June 22, 2024

SPACEY

A COMMERCIAL ROCKET LAUNCHER

Outline

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

Executive Summary

SpaceY is a new competitor in the rocket launch market, aiming to challenge SpaceX.

SpaceX offers launch services starting at $62 million, including reusable first-stage boosters.

Using historical SpaceX data, our predictive models achieved 83.3% accuracy in forecasting booster landing success.

Key Insights:

Payload mass, **launch site**, and **booster version** significantly impact launch success.

Data-driven strategies can optimize SpaceY's operations and bids.

Conclusion: SpaceY can leverage these insights to enhance efficiency and competitiveness in the launch market.

Introduction

- Project Goal: To help SpaceY, a new rocket launch provider, compete with SpaceX by utilizing data analysis and predictive modeling.

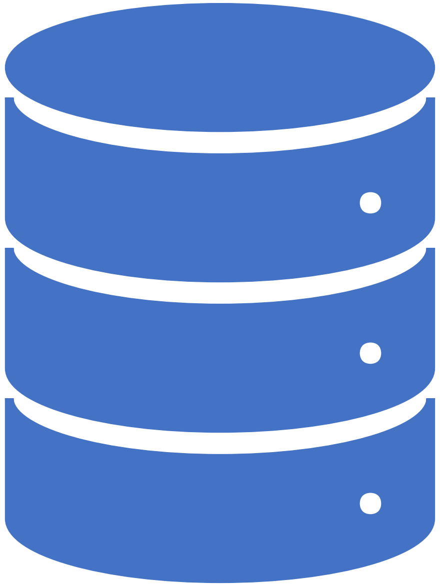
- Objective: Analyze historical launch data from SpaceX to identify key factors influencing first-stage booster landing success.

Section 1

Methodology

Executive Summary of Methodology

- Data Collection: Gathered and preprocessed SpaceX launch data.
- Perform data wrangling: Web scrap Falcon 9 launch records with BeautifulSoup
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models



Data Collection – SpaceX API

- **API**
  - Acquired historical launch data from Open Source REST API for SpaceX
    - Requested and parsed the SpaceX launch data using the GET request
    - Filtered the dataframe to only include Falcon 9 launches
    - Replaced missing payload mass values from classified missions with mean
- [GitHub URL]()

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPa
4	1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	Non
5	2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	Non
6	3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	Non
7	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	Non
8	5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	Non

Data Collection - Scraping

- Web scrap Falcon 9 launch records with BeautifulSoup:
  - Acquired historical launch data from Wikipedia page 'List of Falcon 9 and Falcon Heavy Launches’
  - Parse the table and convert it into a Pandas data frame
- [GitHub URL]

```
[5]: static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
# assign the response to a object
data  = requests.get(static_url).text
```

```
[7]: # Use BeautifulSoup() to create a BeautifulSoup object from a response text content
soup = BeautifulSoup(data)
```

```
df.head(5)
```

	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\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA (COTS)\nNRO	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA (COTS)	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA (CRS)	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Data Wrangling

- Exploratory Analysis:
  - **Launch Sites:** Calculate the number of launches on each site
  - **Orbits:** Calculate the number and occurrence of each orbit
  - **Mission Outcomes:** Calculate the number and occurence of mission outcome of the orbits.
- Create a landing outcome **label** from Outcome column:
  - Training Label: 'Class' for supervised model training.
    - Class = 0 (Unsuccessful Landings)
    - Class = 1 (Successful Landings)
- [GitHub URL]()

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

```
# Apply value_counts on Orbit column
df['Orbit'].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
ES-L1     1
HEO       1
SO        1
GEO       1
Name: Orbit, dtype: int64
```

```
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
False Ocean     2
None ASDS       2
False RTLS      1
Name: Outcome, dtype: int64
```

```
df['Class']=landing_class
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1
7	1

EDA with Data Visualization

- **Data Loading:** The dataset was read into a Pandas dataframe for analysis.
- **Visualization Libraries:** Matplotlib and Seaborn libraries were used for creating visualizations.
- **Plots Created and Their Purpose**:
  - Flight Number vs Payload Mass
    - Purpose: To see if payload mass changes over flights.
    - Reason: Identifies trends in payload capacity improvements.
  - Flight Number vs Launch Site
    - Purpose: To visualize launches at different sites over time.
    - Reason: Shows frequency of site usage and operational importance.
  - Payload vs Launch Site
    - Purpose: To compare payload masses from different sites.
    - Reason: Indicates site capability for handling varying payload sizes.
  - Orbit Type vs Success Rate
    - Purpose: To evaluate mission success for different orbits.
    - Reason: Identifies which orbits have higher or lower success rates.
  - Flight Number vs Orbit Type
    - Purpose: To track target orbits across flights.
    - Reason: Shows changes in mission objectives over time.
  - Payload vs Orbit Type
    - Purpose: To examine payload mass for different orbits.
    - Reason: Identifies suitable payload masses for specific orbits.
  - Year vs Success Rate
    - Purpose: To assess launch success over years.
    - Reason: Reflects improvements or issues in launch success over time.
- GitHub URL

EDA with SQL

- **Loaded Data**: Imported the SpaceX launch dataset into an IBM DB2 database instance for efficient querying and analysis.
- **SQL Queries**: Executed various SQL queries to extract and display key information about the dataset:
    - Launch Sites: Identified and listed the unique launch sites used in the SpaceX missions.
    - Payload Masses: Retrieved payload mass details to understand the distribution and total mass of payloads launched.
    - Booster Versions: Analyzed the different booster versions used in the launches.
    - Mission Outcomes: Listed the outcomes of each mission to assess the success rates.
    - Booster Landings: Examined the details of booster landings to evaluate landing success rates and methods.
- GitHub URL

Build an Interactive Map with Folium

- **Summary of Map Objects:**
  - Markers: Added markers at each launch site to identify their locations.
  - Circles: Created circles around each launch site to visually represent their vicinity.
  - Lines: Drew lines connecting launch sites to nearby cities, railways, highways, and coastlines to analyze their proximities.
- **Explanation of Objects:**
  - Markers: Used to pinpoint the exact locations of the launch sites.
  - Circles: Help to quickly identify the area around each launch site.
  - Lines: Illustrate distances and relationships between launch sites and surrounding infrastructure, which is crucial for logistical analysis.
- **GitHub URL**

Build a Dashboard with Plotly Dash

- **Summary of Plots and Interactions Added:**
  - Dropdown Menu: Allows selection of different launch sites to filter data.
  - Pie Chart: Displays success rates for selected launch sites.
  - Range Slider: Enables selection of payload range to filter data.
  - Scatter Plot: Shows the correlation between payload mass and launch success, color-coded by booster version.
- **Explanation**:
  - Dropdown Menu: Facilitates exploration of specific launch site data.
  - Pie Chart: Provides a quick visual summary of launch success rates.
  - Range Slider: Helps in analyzing data within specific payload ranges.
  - Scatter Plot: Visualizes relationships between payload mass and launch success, enhancing data-driven decision making.
- GitHub URL

Predictive Analysis (Classification)

- **Model Development Summary:**
  - Data Preparation: Collected and cleaned historical launch data from SpaceX API and Wikipedia.
  - Feature Engineering: Created and selected relevant features, applied one-hot encoding.
  - Model Building: Built classification models (Logistic Regression, Decision Tree, SVM, KNN) and tuned hyperparameters.
  - Evaluation: Compared models using accuracy and F1-score, selected the best model.
  - Improvement: Fine-tuned parameters and enhanced features.
  - Final Model: Logistic Regression with **83.3%** accuracy.
- **GitHub URL**

Results

Exploratory data analysis results

Interactive analytics demo in screenshots

Predictive analysis results

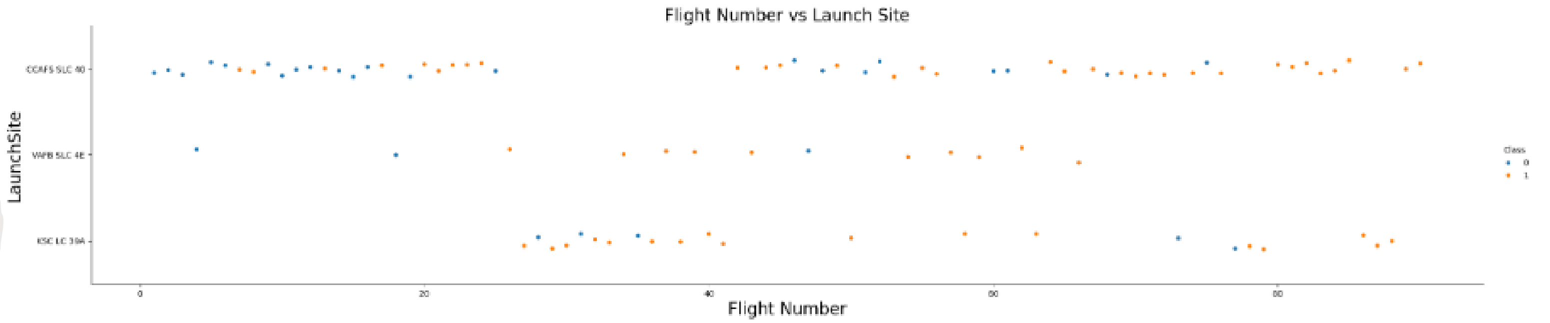
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Key Observations:

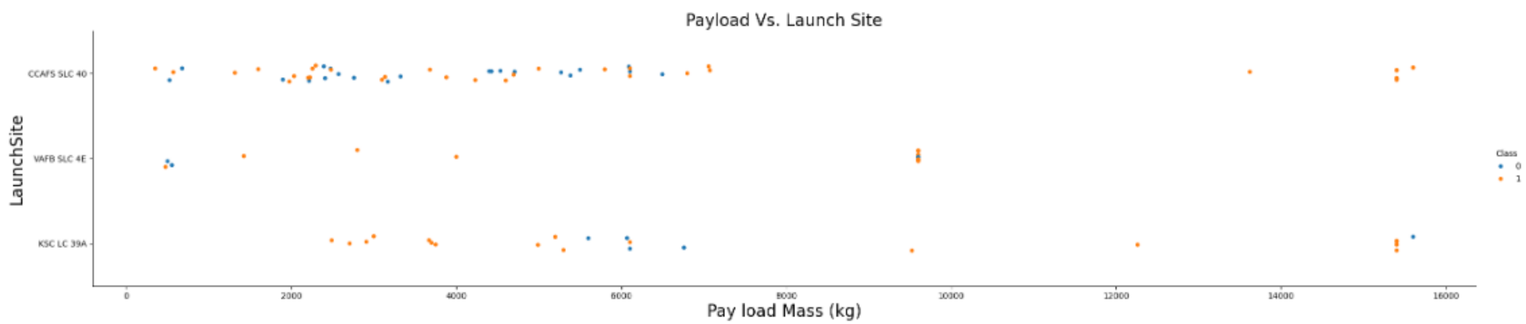
- **Consistent Launch Sites:** The three main launch sites (CCAFS SLC 40, VAFB SLC 4E, and KSC LC 39A) are consistently used throughout the timeline.
- **Success Distribution:** There is a noticeable pattern in the distribution of successful launches across the different sites, with some sites showing higher success rates over time.
- **Trends Over Time:** The data suggests improvements in launch success rates as the flight numbers increase, potentially indicating improvements in technology and processes



Payload vs. Launch Site

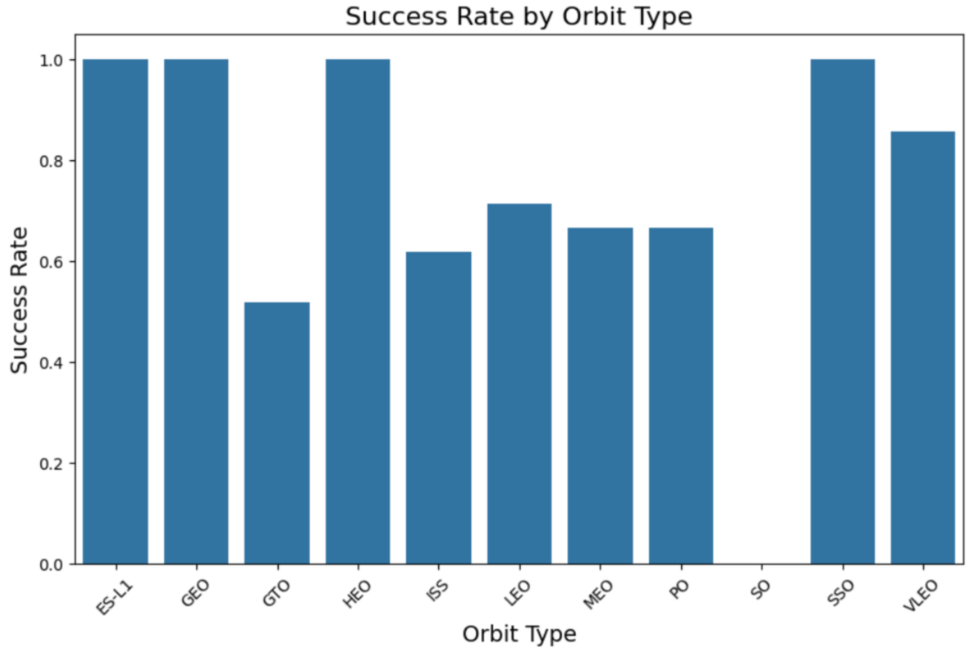
- **Observation**: There is no clear correlation between the payload mass and the success of the launch. However, KSC LC 39A appears to handle larger payloads more frequently.

- **Insight**: This information helps in understanding the performance and capacity of different launch sites relative to the payload mass they handle.



Success Rate vs. Orbit Type

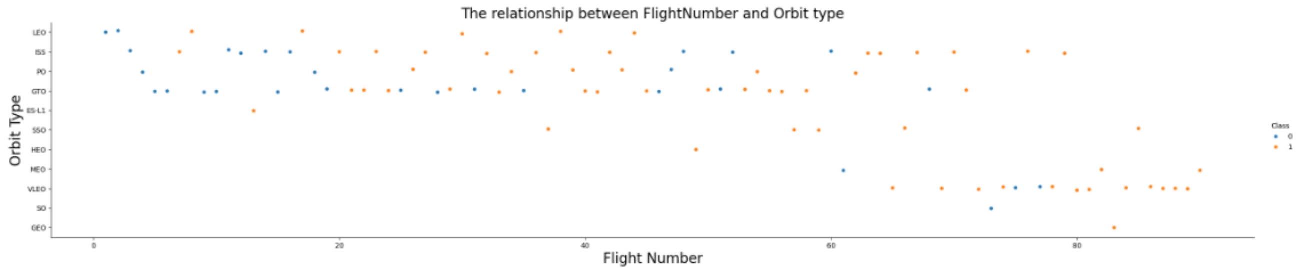
- **Observation**: Certain Orbit types like ES-L1, GEO, and SSO show a 100% success rate, indicating high reliability, whereas others like GTO have lower success rates, highlighting potential challenges.



Flight Number vs. Orbit Type

Observation:

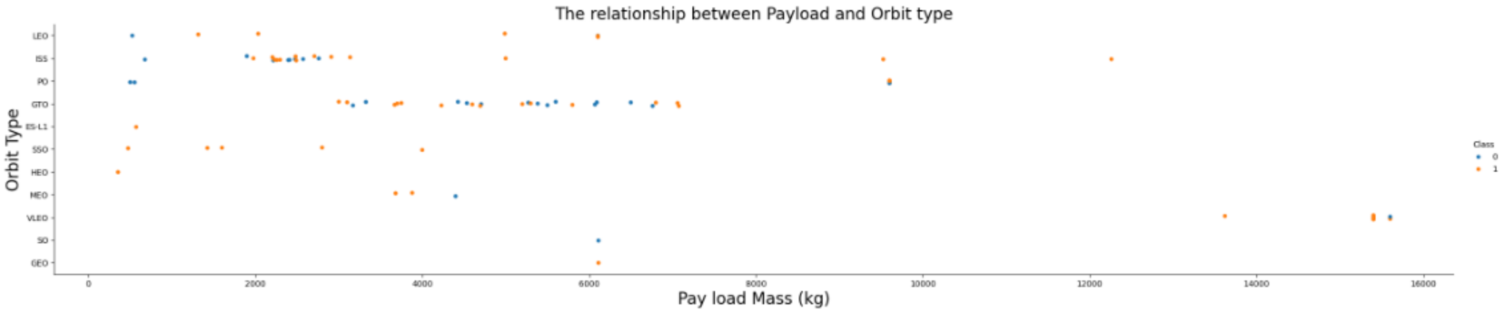
- Flight Number: As the flight number increases, it suggests advancements in technology and experience over time.
- Orbit Type: Different orbit types are scattered across the plot, indicating the variety of missions targeting different orbital paths.
- Class: The distribution of blue and orange dots shows the success rate across different orbit types and flight numbers.



Payload vs. Orbit Type

Observation:

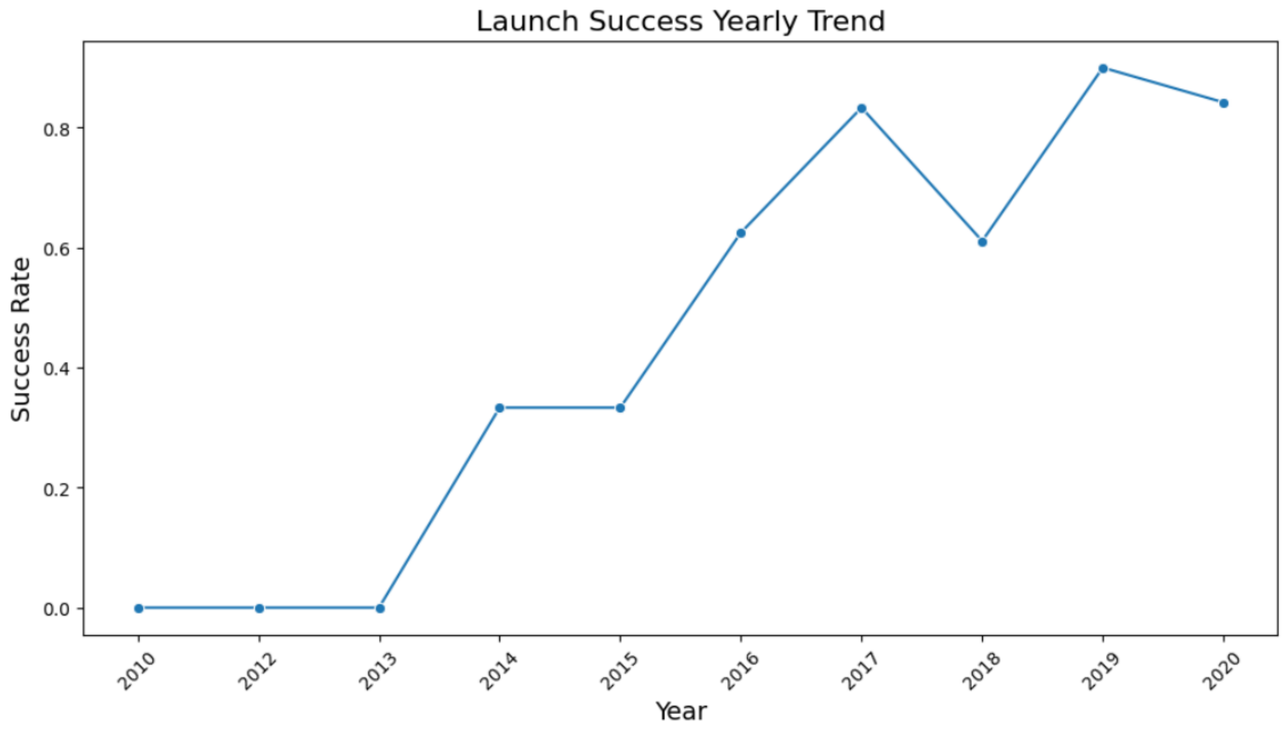
- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

Observation:

- This line chart showing improvements or declines over time. It helps identify any long-term trends or patterns in launch success.
- We can observe that the success rate since 2013 kept increasing till 2020.



All Launch Site Names (SQL)

Explanation:

- The query retrieves distinct launch sites from the SPACEXTABLE dataset. This helps identify the unique locations from which the SpaceX launches have occurred.

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE "CCA%" Limit 5
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_O
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (pa
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 (pa
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No

- These records show the launches from the Cape Canaveral Air Force Station (CCAFS) LC-40 launch site

Total Payload Mass

- This query calculates the sum of payload masses for all launches where the customer is NASA (CRS). This sum represents the total mass of payloads that NASA (CRS) has sent into space using the SpaceX launch services, which amounts to 45,596 kilograms.

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total payload mass by NASA(CRS)" FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
```

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

Total payload mass by NASA(CRS)
45596

Average Payload Mass by F9 v1.1

- This query calculates the average payload mass carried by the Falcon 9 booster version 1.1. It provides an insight into the average performance of a specific booster version in terms of payload capacity.

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "AVG payload mass by booster version F9 v1.1" FROM SPACEXTABLE WHERE Booster_Version LIKE "F9 v1.1%'
```

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

AVG payload mass by booster version F9 v1.1
2534.6666666666665

First Successful Ground Landing Date

- 2015-12-22 is the date when the first succesful landing outcome in ground pad was acheived.

```
%sql SELECT MIN(Date) AS "Date" FROM SPACEXTABLE WHERE Landing_Outcome LIKE"Success%"
```

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

Date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query lists the booster versions that have successfully landed on a drone ship while carrying payloads between 4000 and 6000 kg, indicating the boosters' ability to handle significant payloads and achieve successful landings.

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)' AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS_
```

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

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- Present your query result with a short explanation here

```
%sql SELECT Mission_Outcome, COUNT(*) as Outcome_Count FROM SPACEXTABLE GROUP BY Mission_Outcome
```

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

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

Boosters Carried Maximum Payload

- This query retrieves the booster versions that carried the maximum payload mass. It highlights the boosters that achieved the highest payload capacities in the dataset.

```
%sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
```

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

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

- This query retrieves records of launches where the landing outcome was "Failure (drone ship)"in the months in year 2015. It highlights the months, booster versions, and launch sites associated with these failures.

```
%sql SELECT strftime('%m', Date) as Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE Landing_Outcome = 'Failur
```

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

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

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

- This query counts "Failure (drone ship)", "Success (ground pad)") from the SPACEXTABLE between the dates 2010-06-04 and 2017-03-20. It groups the results by Landing_Outcome and sorts them in descending order of count.

Rank 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.

```
%sql SELECT Landing_Outcome, COUNT(*) as Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY Landing_
```

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

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

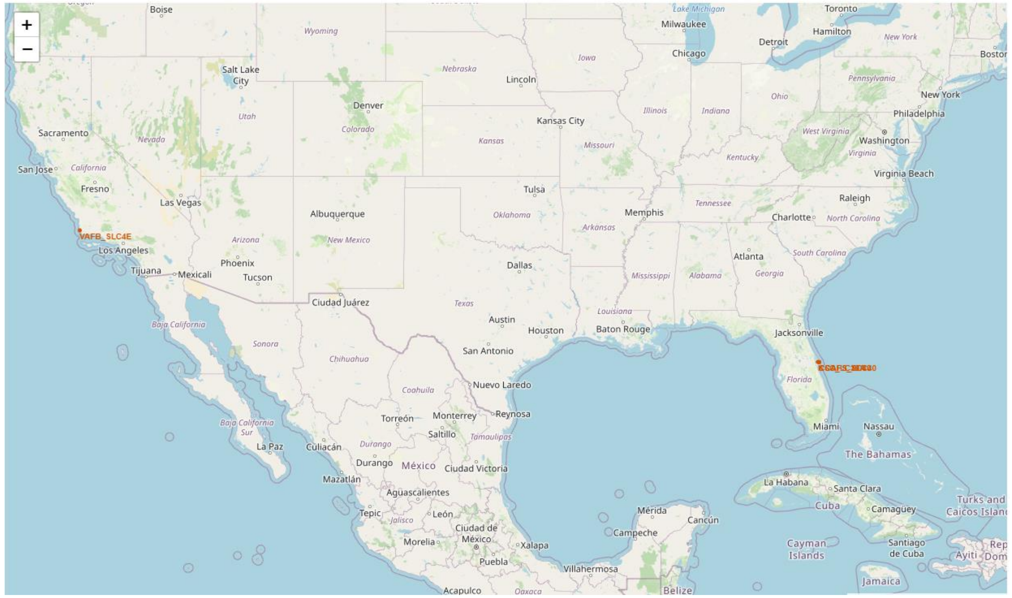
Section 3

Launch Sites Proximities Analysis

Launch Sites Locations Analysis with Folium

Explanation of Findings:

- Launch sites are strategically placed near coasts to minimize the risk to populated areas in case of launch failures, as well as to facilitate easier transport of rockets and equipment via sea routes.
- Proximity to the Equator is advantageous for launches into equatorial orbits due to the higher rotational speed of the Earth, but it is not necessary for all types of missions. The launch sites on the map are chosen for other logistical and safety reasons.



Launch Sites Locations Analysis with Folium

Findings:

- KSC LC-39A has a higher success rate compared to CCAFS SLC-40.
- Both sites are coastal, which is typical for space launches to reduce risks and facilitate logistics.

KSC LC-39A

CCAFS SLC-40

Launch Sites Locations Analysis with Folium

CCAFS_SLC40 Proximity to Railways: 1.26 KM

CCAFS_SLC40 Proximity to Highways: 0.08 KM

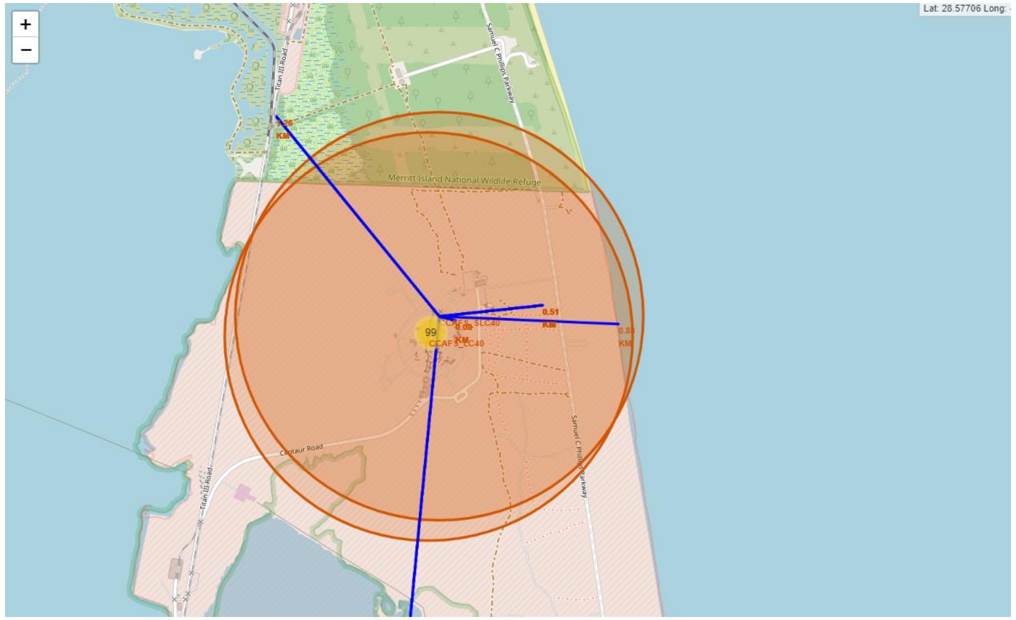
CCAFS_SLC40 Proximity to Coastline: 0.88 KM

CCAFS_SLC40 Proximity to Cities: 51.53 KM

Findings:

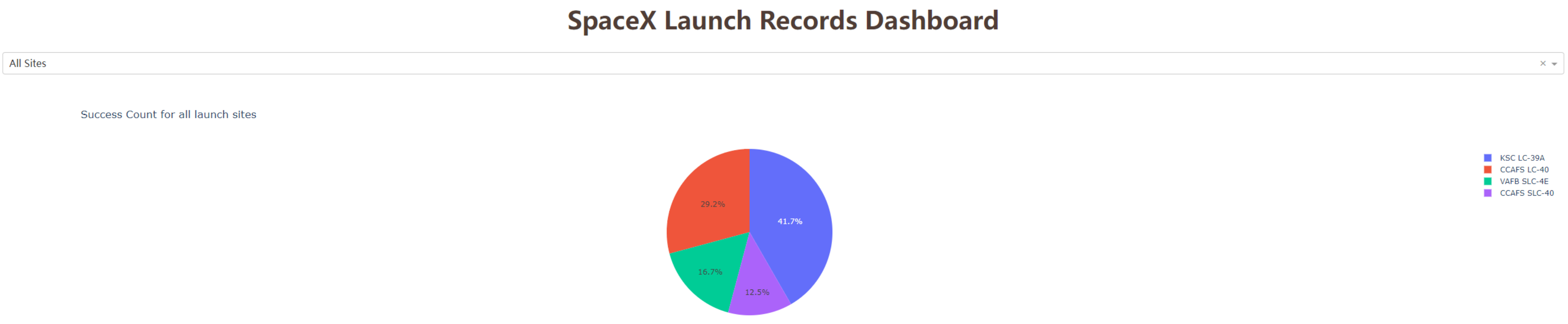
- The launch site's strategic locations near railways and highways ensure efficient logistical support while the proximity to the coastline provides safety and recovery benefits. The distance from cities helps mitigate risk and disturbance to the population. These proximities reflect the thoughtful planning and safety considerations taken into account when selecting and operating launch sites.

CCAFS_SLC40



Section 4

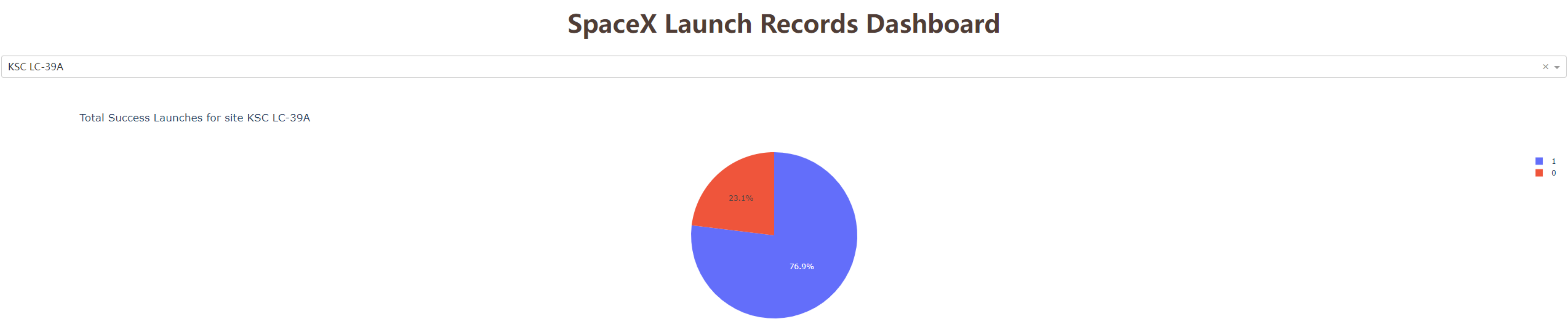
Build a Dashboard with Plotly Dash



Interactive Dashboard with Ploty Dash

Findings:

- **Success Rates:**
  - The pie chart indicates that the highest proportion of successful launches occurred at KSC LC-39A, accounting for 41.7% of the total successful launches.
  - CCAFS LC-40 follows with 29.2%, VAFB SLC-4E with 16.7%, and CCAFS SLC-40 with 12.5%.
- **Launch Site Comparison:**
  - The data suggests that KSC LC-39A is the most successful launch site among the ones listed, followed by CCAFS LC-40. This can be useful for understanding which sites have higher reliability and success rates.



Interactive Dashboard with Ploty Dash

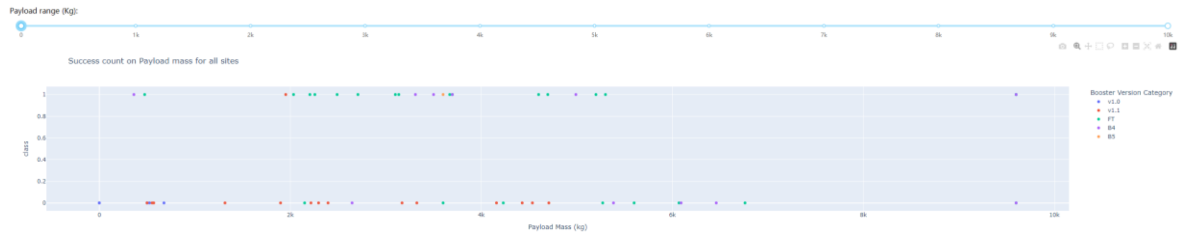
Finding:

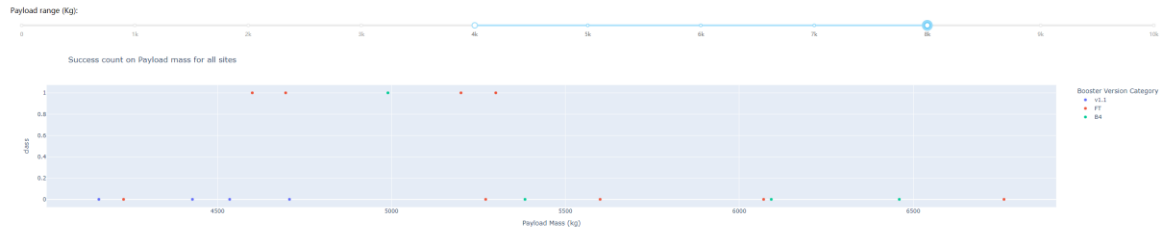
- Success Rate at KSC LC-39A:The pie chart indicates that the majority of launches at KSC LC-39A have been successful, accounting for 76.9% of the total launches from this site.
- A smaller proportion, 23.1%, represents unsuccessful launches.

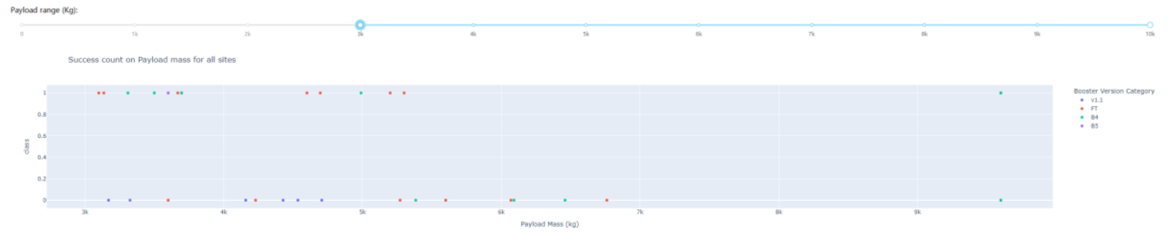
Interactive Dashboard with Ploty Dash

Finding:

- The scatter plots provide a clear visualization of the relationship between payload mass, booster versions, and the success rate of launches. The findings suggest that newer booster versions (F9 FT and F9 B5) have higher success rates, and most successful launches are concentrated in the lower to mid payload mass ranges. This information is crucial for understanding the performance and reliability of different booster versions and their capability to handle varying payload masses.





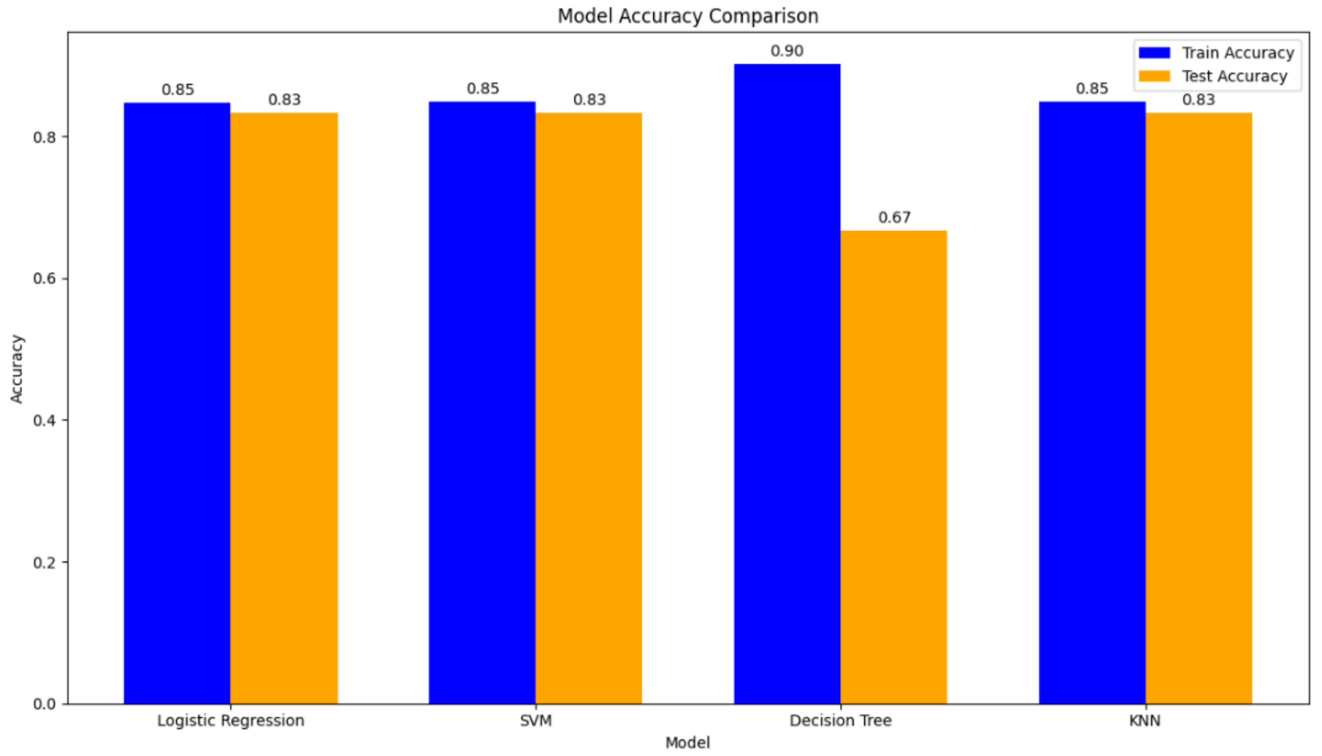


Section 5

Predictive Analysis (Classification)

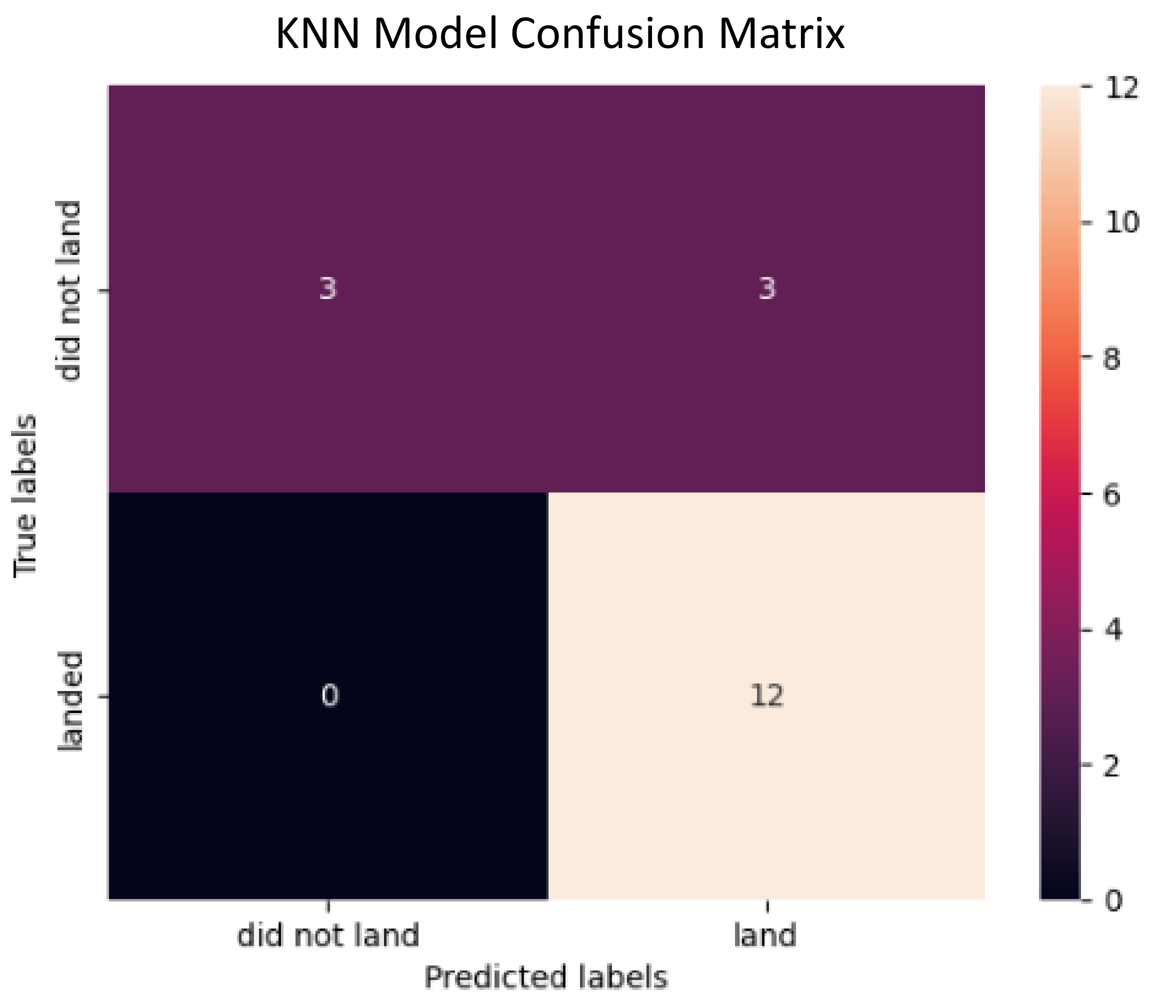
Classification Accuracy

- These models have similar performances with both train and test accuracies around 0.83-0.85, indicating they generalize well and perform reliably on new data. However, the **Decision Tree** model, despite having the highest train accuracy, shows signs of overfitting due to its significantly lower test accuracy, making it less reliable on new data.
- In summary, **Logistic Regression**, **SVM**, and **KNN** are the best models due to their balanced and consistent performance on both training and test datasets.



Confusion Matrix

- The KNN model shows a good overall accuracy of 0.8333, with high recall (1.0) indicating it correctly identifies all the positive instances (landed). However, it has a few false positives, as shown by the precision (0.8). The F1 Score of approximately 0.8889 reflects a balanced performance considering both precision and recall.



Conclusions

Prediction Accuracy: Using the models developed in this report, SpaceY can predict the success of the first stage booster landing with 83.3% accuracy. This allows for more informed decision-making regarding launch bids.

Competitive Advantage for SpaceY: By understanding these key factors, SpaceY can leverage these insights to offer competitive pricing and improve the likelihood of successful bids, enhancing its market position.

Future Improvements

Model Refinement: Continuously updating and refining models with new data will enhance prediction accuracy.

Data Integration: Incorporating more launch data and outcomes will provide deeper insights and improve model robustness.

Subdividing Models: Developing specialized models to predict specific outcomes (e.g., whether SpaceX will attempt or succeed in landing) can provide more nuanced forecasts.

Appendix

- Notebooks links to recreate dataset, analysis, and models:

  1. Collecting the data: https://github.com/hx2163/IBM-data-science-professional-certificate/blob/main/10.Applied%20Data%20Science%20Capstone/SpaceX%20rocket%20launch%20analysis/SpaceX_Data_Collection_API.ipynb
  2. Web scraping: https://github.com/hx2163/IBM-data-science-professional-certificate/blob/main/10.Applied%20Data%20Science%20Capstone/SpaceX%20rocket%20launch%20analysis/web_scraping.ipynb
  3. Data wrangling: https://github.com/hx2163/IBM-data-science-professional-certificate/blob/main/10.Applied%20Data%20Science%20Capstone/SpaceX%20rocket%20launch%20analysis/SpaceX_Data_wrangling.ipynb
  4. Exploring and Preparing Data: https://github.com/hx2163/IBM-data-science-professional-certificate/blob/main/10.Applied%20Data%20Science%20Capstone/SpaceX%20rocket%20launch%20analysis/SpaceX_Exploring_Preparing_Data.ipynb
  5. Launch Sites Locations Analysis with Folium: https://github.com/hx2163/IBM-data-science-professional-certificate/blob/main/10.Applied%20Data%20Science%20Capstone/SpaceX%20rocket%20launch%20analysis/Sites_Locations_Analysis_with_Folium.ipynb
  6. SQL for SpaceX dataset: https://github.com/hx2163/IBM-data-science-professional-certificate/blob/main/10.Applied%20Data%20Science%20Capstone/SpaceX%20rocket%20launch%20analysis/SQL_for_SpaceX_DataSet.ipynb
  7. Machine Learning Prediction: https://github.com/hx2163/IBM-data-science-professional-certificate/blob/main/10.Applied%20Data%20Science%20Capstone/SpaceX%20rocket%20launch%20analysis/SpaceX_ML_Prediction.ipynb

- References

  1. https://api.spacexdata.com/v4/launches/past
  2. https://www.spacex.com/vehicles/falcon-9/

Thank you!