DOVUD ASADOV
29/02/24

SPACEY

AGENDA

1. Executive Summary
2. Introduction
3. Methodology
4. Visualization
5. Maps with Folium
6. Plotly Dash board
7. Predictive Analytics
8. Conclusion

Executive Summary

Methodologies

- Dual methodology: API integration and web scraping for data collection.
- Python data manipulation for dataset processing and cleansing.
- SQL queries for extracting relevant information.
- Data visualization and trend analysis for early insights.
- Supervised machine learning models for predicting landing event success.

Results

- Identified patterns and correlations influencing landing event success.
- Developed and trained a predictive model with 83% accuracy.
- Accurately forecasted the probability of a successful landing event.

Improvement

- Dual methodology: API integration and web scraping.
- Python data manipulation for dataset processing.
- SQL queries for relevant information extraction.
- Data visualization and trend analysis for insights.
- Supervised machine learning models for accurate predictions.

Introduction

Background

SpaceX, short for Space Exploration Technologies Corp., is an American aerospace manufacturer and space transportation company founded by Elon Musk in 2002. The company's primary goal is to revolutionize space technology and make space exploration more accessible and sustainable.

One of SpaceX's key achievements is the development of the Falcon 1, Falcon 9, and Falcon Heavy rockets, which are designed with reusability in mind. By successfully recovering and reusing the first stages of these rockets, SpaceX has made significant strides in reducing the cost of space travel. The company has demonstrated the ability to land and refurbish rocket boosters, a feat that was once thought to be impossible.

In addition to its focus on reusable rockets, SpaceX is known for its ambitious goal of colonizing Mars. The company's Starship spacecraft, currently under development, is intended to transport humans and cargo to the Red Planet. Elon Musk envisions establishing a self-sustaining colony on Mars in the future, making humanity a multi-planetary species.

Explore

In order to evaluate the efficiency and cost-effectiveness of SpaceX's pioneering approach, our project conducted an in-depth analysis of the success rate of these retrieval events. We recognized that the success of the first-phase rocket retrieval serves as a valuable metric for assessing the overall effectiveness of SpaceX's reusable rocket program. By closely examining the data associated with these events, we aimed to gain insights into the factors that contribute to successful retrievals and identify areas for improvement.

METHODOLOGY

Methodology

Steps

- Gather data using a combination of SpaceX REST API integration and web scraping techniques.
- Perform data wrangling tasks, including filtering data, handling missing values, and applying one-hot encoding, to prepare the dataset for analysis and modeling.
- Explore the dataset through exploratory data analysis (EDA) using SQL queries and employ data visualization techniques to gain insights and identify patterns.
- Utilize tools like Folium and Plotly dash to create interactive and visually appealing visualizations of the data.
- Develop predictive models using classification algorithms to forecast the outcomes of rocket landings.
- Fine-tune and evaluate the models by adjusting parameters and employing performance metrics to identify the best-performing model.
- Assess the accuracy and effectiveness of the final model in predicting landing outcomes.
- Provide recommendations and insights based on the analysis and modeling results to enhance decision-making within the space industry.

Data Collection-SpaceX API

1.Define Auxiliary Functions to Parse the Data:

- Create auxiliary functions that handle the parsing of specific data elements from the API response.

2.Retrieve Data from the REST API using the GET Method:

- Utilize the appropriate HTTP library to send a GET request to the REST API endpoint.
- Include any necessary query parameters or headers required to authenticate or access the data.

3. Parse the Data with the Auxiliary Functions:

- Extract the relevant data from the API response using the auxiliary functions defined earlier.
- Apply any necessary transformations or data manipulations to ensure the extracted data is in the desired format.

4.Store the Data in Pandas DataFrame:

- Create an Pandas DataFrame with the appropriate column names to store the parsed data.
- Iterate over the parsed data and append each row to the DataFrame using Pandas' DataFrame.append() or DataFrame.loc[] method.

Data Collection-Web Scraping

1. Using the requests.get Method to Download Page Code:
   - Use the get method from the requests library to send a GET request to the desired URL and store the response in a variable.
   - Access the content of the response using the .text attribute.
2. **Creating a BeautifulSoup Object to Manipulate the HTML Text:**
   - Import the BeautifulSoup class from the bs4 (BeautifulSoup) library.
   - Create a BeautifulSoup object by passing the HTML text from the response as the first argument and specifying the HTML parser (e.g., 'html.parser').
3. **Filtering the Desired Tables using Soup Manipulation Methods:**
   - Inspect the HTML structure of the page to identify the specific tables you want to extract.
   - Use BeautifulSoup's methods to navigate and filter the HTML structure to locate the desired tables. For example, you can use find, find_all, or CSS selectors.
4. **Converting the Data from HTML to Pandas DataFrame Format:**
   - Create an empty Pandas DataFrame with the desired column names.
   - Iterate over the filtered HTML tables and extract the data for each row.
   - Append each row of data to the Pandas DataFrame using the .loc or .append method, depending on your preference.

Data Wrangling

1. **EDA and Determining Data Labels:**
   - Perform exploratory data analysis (EDA) on the dataset to gain insights into its structure, distributions, and potential relationships.
   - Identify the relevant variables for analysis, such as launch sites, orbits, mission outcomes, and landing outcomes.
2. **Calculation of Launches for Each Site:**
   - Group the data by launch sites and count the number of launches for each site.
   - Calculate the occurrence or frequency of launches at each site.
3. **Calculation of Occurrence of Orbits:**
   - Group the data by orbit types and count the number of occurrences for each orbit type.
   - Calculate the frequency or distribution of different orbit types.
4. **Calculation of Occurrence of Mission Outcome per Orbit Type:**
   - Group the data by orbit types and mission outcomes.
   - Calculate the occurrence or frequency of each mission outcome within each orbit type.
5. **Creating Binary Target Column:**
   - Based on the landing outcomes, create a binary target column that represents specific conditions such as True for landing in the ocean, RTLS (Return to Launch Site), or ASDS (Autonomous Spaceport Drone Ship), and False otherwise.

EDA with Visualization

Data Visualization:

- Utilize various visualization techniques to explore the dataset visually.
- Create plots such as histograms, bar plots, scatter plots, box plots, or line plots to analyze the distribution, relationships, and trends in the data.
- Use appropriate visualization libraries: matplotlib and seaborn

Analyze Relationships:

- Explore the relationships between variables using techniques like scatter plots, correlation matrices, or heatmaps.
- Identify any correlations or dependencies between variables.

Extract Insights:

- Interpret the visualizations and draw meaningful insights from the data.
- Identify notable patterns, trends, or outliers within the dataset.

Charts:

- Flight number vs. Launch Site
- Flight number vs. Orbit type
- Payload and Launch Site
- Success rate of each orbit type

EDA with SQL

Display:

- Number of unique launch sites
- 5 records where launch site begins with “CCA”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1

Lists:

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster version which have carried the max payload
- Failed landing outcomes on drone ship their booster version and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20

Map with Folium

Marking Launch Sites on the Folium Map:

- Added blue circle at NASA Johnson Space Center's coordiate with a popup label showing its name using its latitude and longitude coordinates
- Added red circle at all launch sites coordinates with a popup label showing its name using its latitude and longitude coordinates

Identifying High Success Rate Launch Sites:

- Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch have high success rates

Calculating Distances to Proximities:

- Added color lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway, and city

Dashboard with Plotly

Dropdown List with launch

• Allow user to select all launch sites or a certain launch site

Slider of Payload Mass Range

• Allow user to select payload mass range

Pie Chart Showing Successful Launches

• Allow user to see successful and unsuccessful launches as a percent of the total

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

• Allow user to see the correlation between Payload and Launch Success

Predictive Analytics

- Create NumPy array from the Class column
- Standardize the data with StandardScaler. Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- Apply GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier()) Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model using Jaccard_Score, F1_Score and Accuracy

RESULTS

Results Summary

Exploratory Data Analysis:

- **Launch Success Trend:** The analysis reveals a positive trend in launch success rates over time, indicating advancements in technology and operational practices.
- **KSC LC-39A Performance:** KSC LC-39A stands out with the highest success rate among landing sites, showcasing its reliability and operational excellence.
- **Orbital Success Rates:** Specific orbits, including ES-L1, GEO, HEO, and SSO, have achieved a perfect 100% success rate, highlighting the precision and effectiveness of missions to these destinations.

Visual Analytics:

- **Geographical Location:** Most launch sites are strategically located near the equator, leveraging the Earth's rotational speed for efficiency. Additionally, all sites are situated near coastal areas, facilitating logistical operations and ensuring a safe launch environment.
- **Safety Considerations:** Launch sites are carefully positioned to minimize risks to populated areas or critical infrastructure in the event of a launch failure, while still maintaining accessibility for operational needs.

Predictive Analytics:

- **Best Predictive Model:** The decision tree model emerges as the top-performing predictive model for the dataset, showcasing its effectiveness in predicting launch outcomes based on the available features.

Flight Numbers vs. Payload Mass

Correlation between Flight Numbers and Landing Success:**

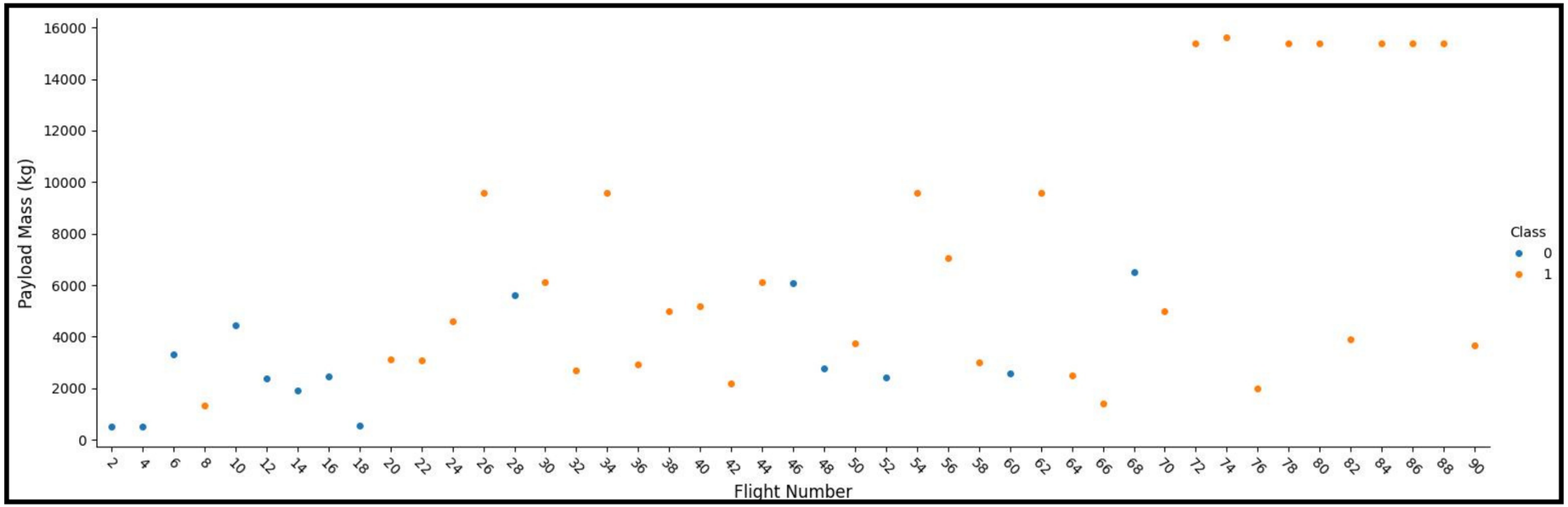
- Analysis indicates a positive correlation between higher flight numbers and a higher likelihood of successful first-stage landings. This trend suggests that experience and iterative improvements play a significant role in enhancing landing success rates over time.

Relationship between Payload Mass and Landing Success:

- An inverse relationship is observed between payload mass and the success of first-stage landings. Heavier payloads are less likely to result in successful landings, highlighting the challenges associated with heavier loads and the need for precise landing capabilities.

Variability in Success Rates across Launch Sites:

- Launch sites exhibit varying success rates, indicating differences in operational effectiveness and environmental factors. Specifically, CCAFS LC-40 has a 60% success rate, while KSC LC-39A and VAFB SLC 4E boast higher success rates of 77%. These differences underscore the importance of site-specific considerations in achieving successful first-stage landings.



Flight Number vs. Launch Site

Success Rate Trends Over Time:

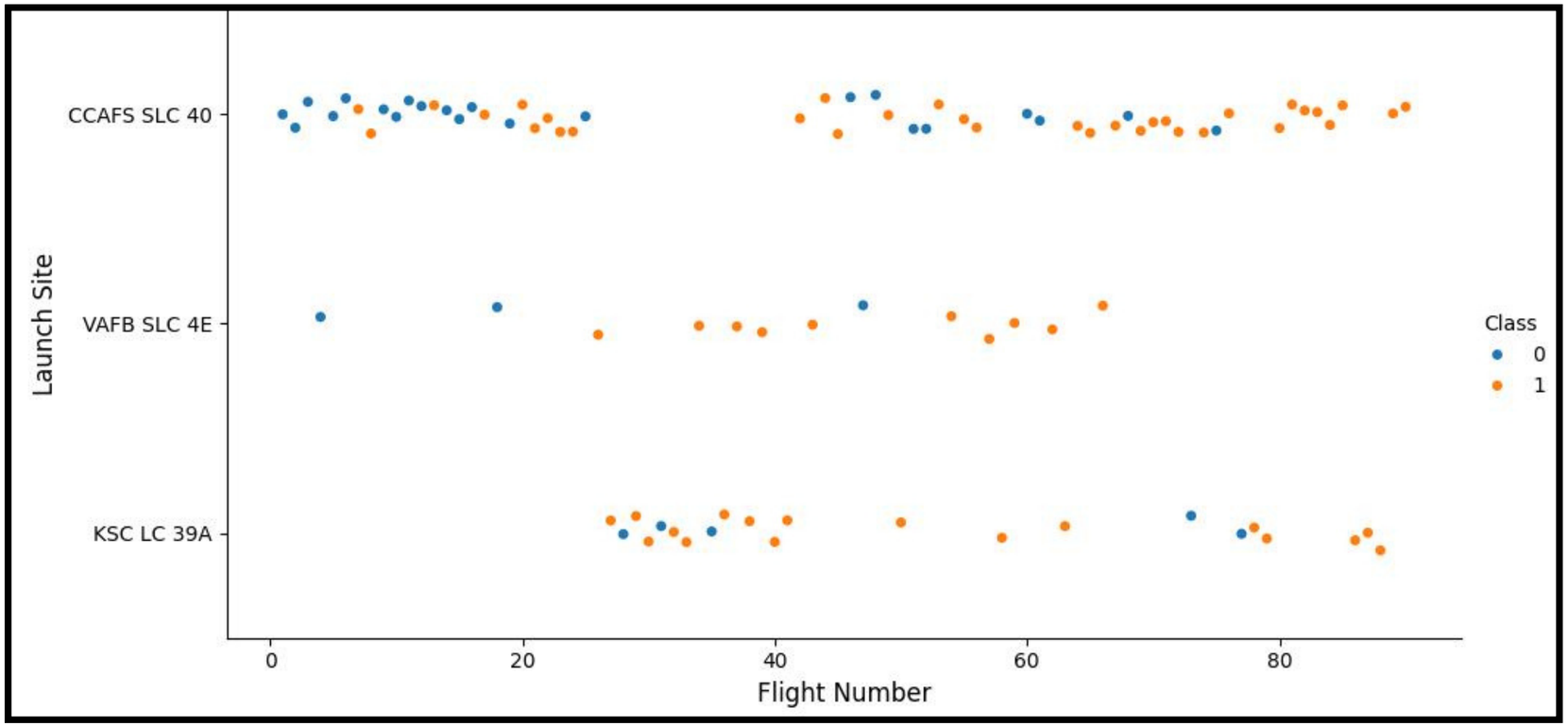
- Earlier flights had lower success rates (blue = fail), while later flights showed higher success rates (orange = success), indicating an improvement in success rates over time.
- About half of the launches were from CCAFS SLC 40.

Success Rates by Launch Site:

- VAFB SLC 4E and KSC LC 39A had higher success rates, suggesting operational advantages or superior infrastructure.

Inference on New Launch Success:

- Newer launches have a higher success rate, reflecting ongoing advancements in space launch technologies and practices.



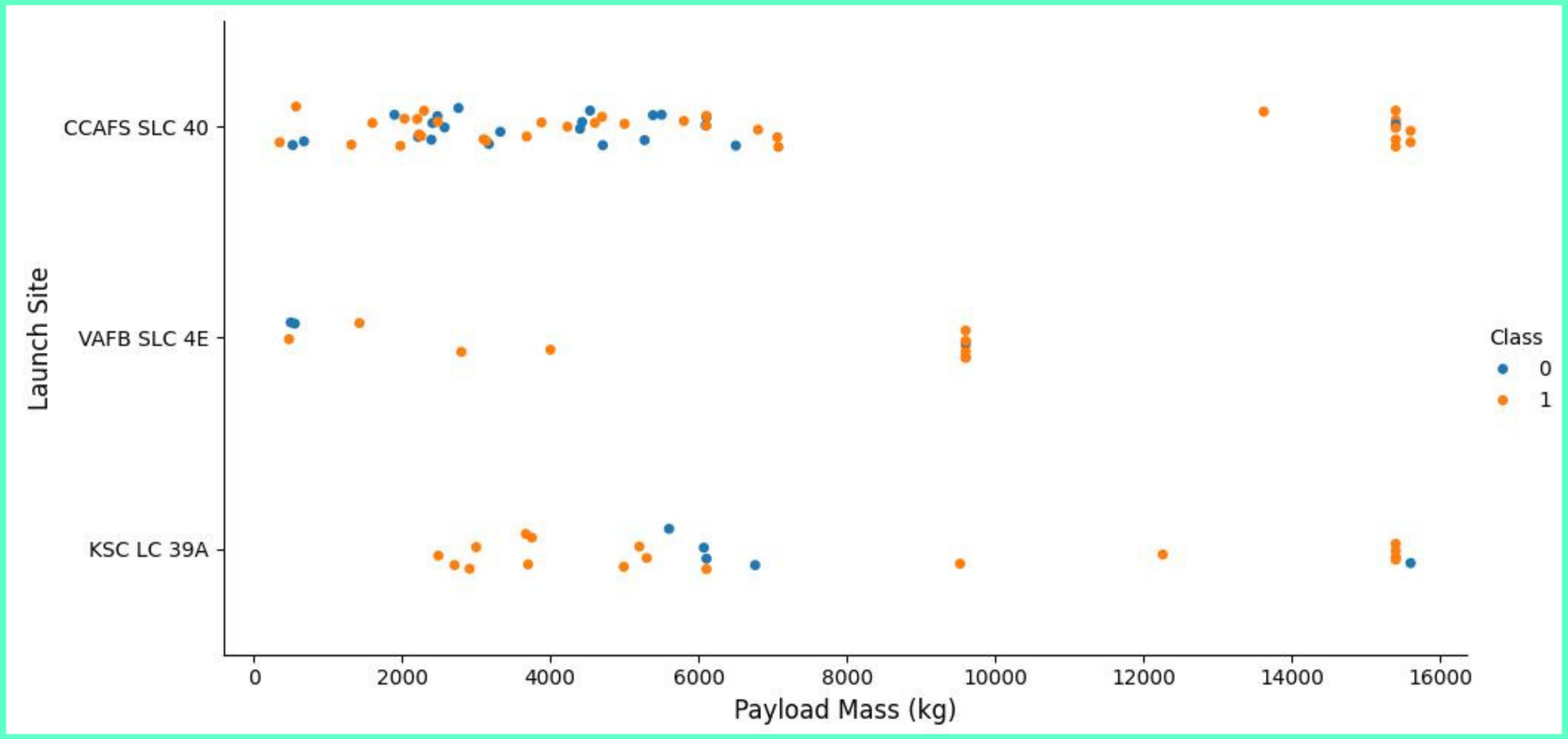
Payload vs. Launch Site

Payload Mass and Success Rate:

- Generally, a higher payload mass correlates with a higher success rate.
- Most launches with a payload greater than 7,000 kg were successful.

Success Rates by Launch Site:

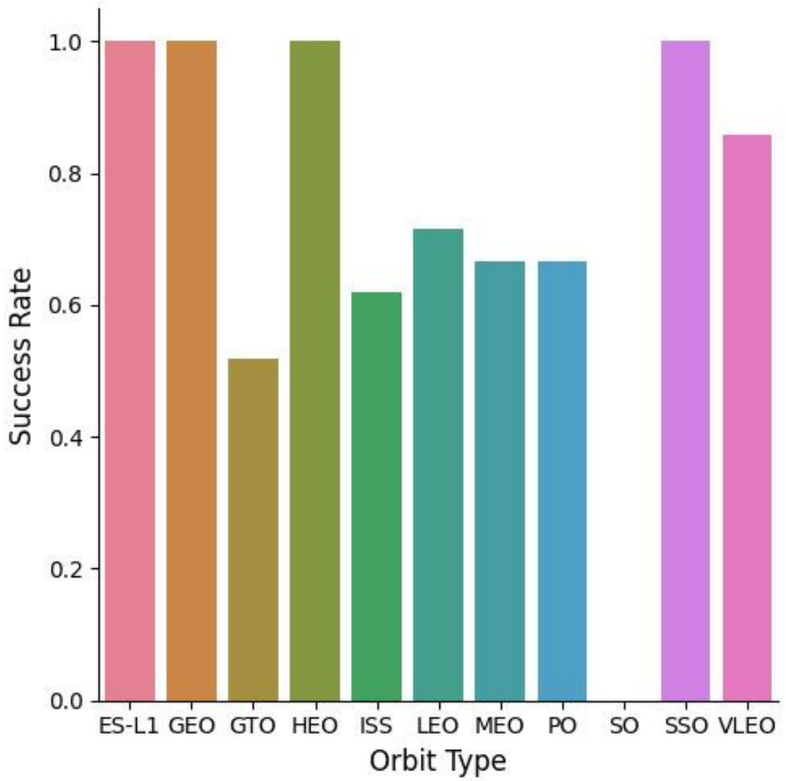
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg.
- VAFB SLC 4E has not launched anything greater than approximately 10,000 kg.



Success Rate by Orbit

Orbital Destination Success Rates:

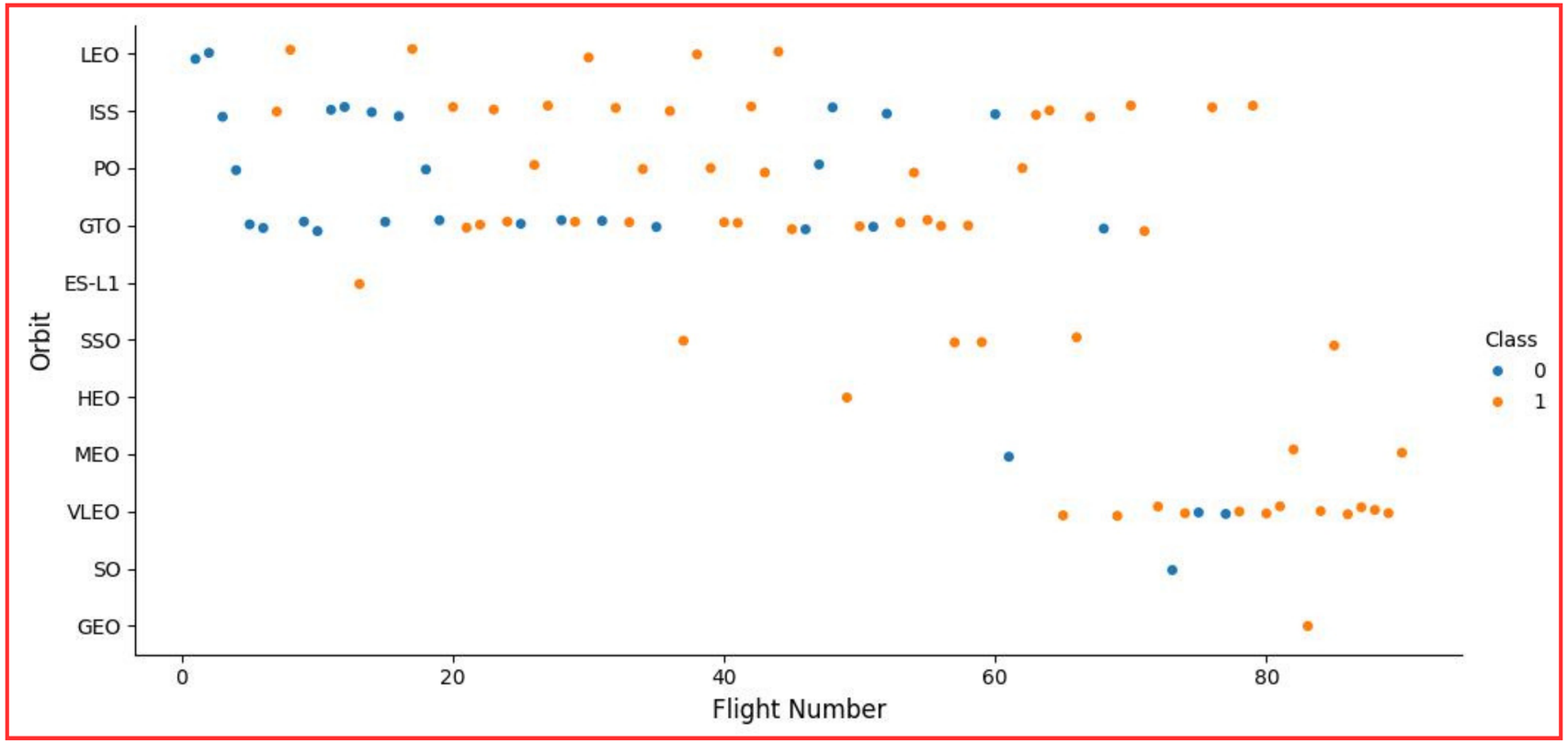
- ES-L1, GEO, HEO, and SSO orbits achieved a 100% success rate.
- Orbits such as GTO, ISS, LEO, MEO, and PO had success rates ranging from 50% to 80%.
- SO orbit had a 0% success rate.



Flight Number vs. Orbit

Success Rate and Number of Flights:

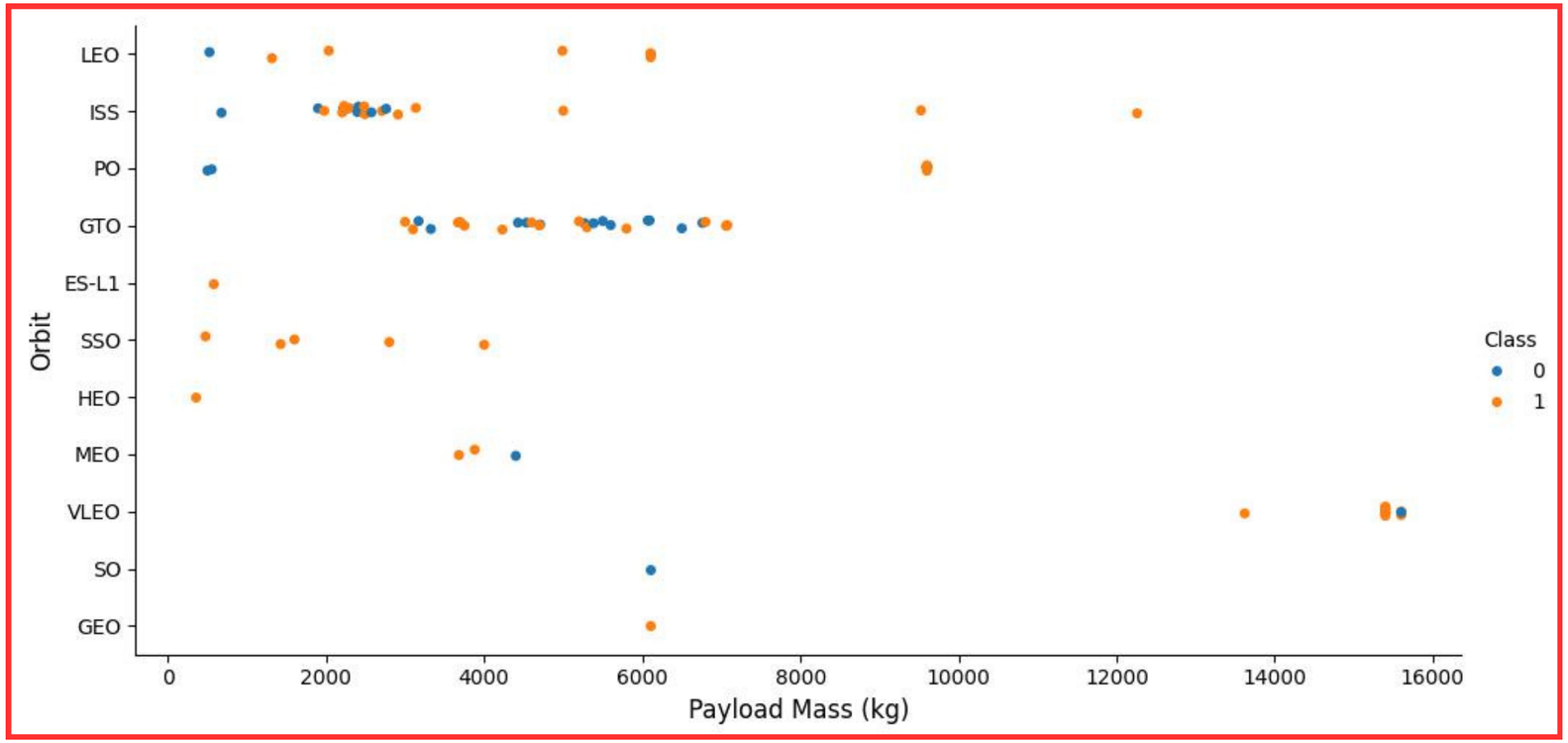
- Generally, the success rate increases with the number of flights for each orbit.
- This relationship is particularly strong for the LEO orbit.
- However, the GTO orbit does not follow this trend.



Payload vs. Orbit

Payload Mass and Orbit Suitability:

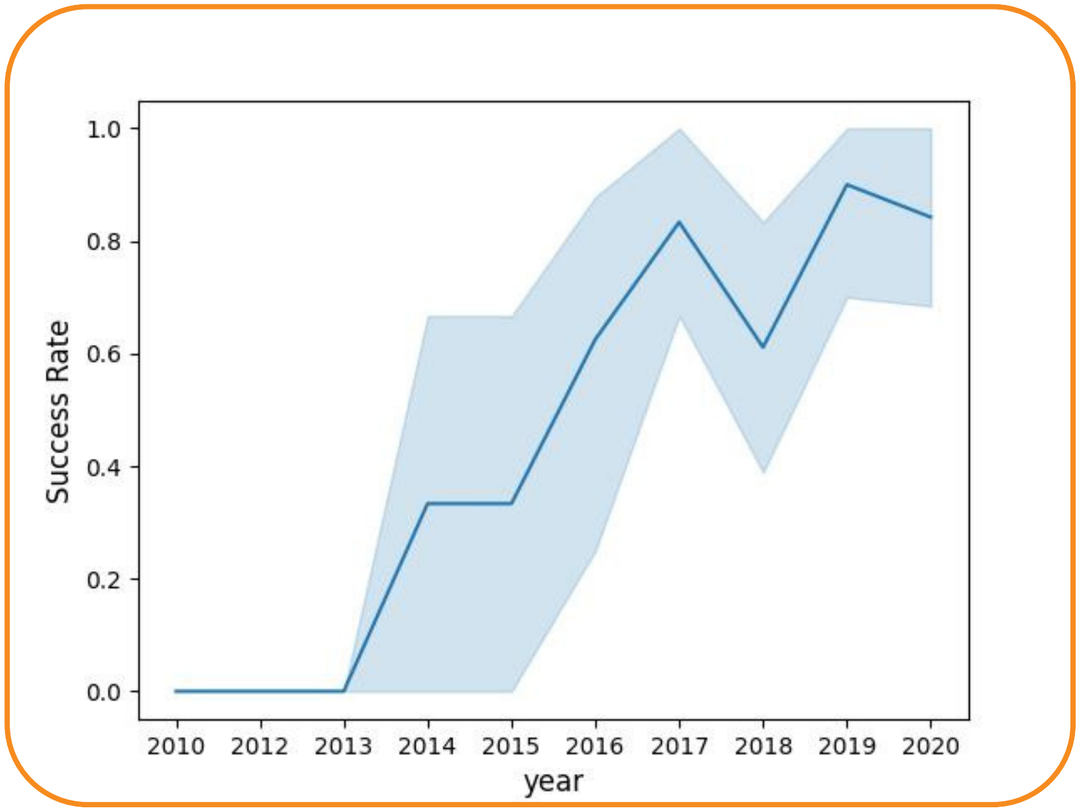
- Heavy payloads are better suited for LEO, ISS, and PO orbits.
- The GTO orbit has mixed success with heavier payloads.



Launch Success over Time

Success Rate Trends Over Time

- The success rate showed a positive trend from 2013 to 2017, indicating improvements in technology, processes, and operational efficiency during this period.
- However, there was a slight decline in success rate from 2017 to 2018, suggesting potential challenges or setbacks faced during that period.
- The success rate improved again from 2018 to 2019, indicating a recovery or further advancements.
- However, another decrease was observed from 2019 to 2020, highlighting potential difficulties or changes in the space industry landscape during that time.
- Despite these fluctuations, the overall trend since 2013 has been positive, indicating a general improvement in launch success rates over the years.



Launch Site Information

Launch Site Names
- CCAFS LC-40
- CCAFS SLC-40
- KSC LC-39A
- VAFB SLC-4E

```
Done.
```

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

Payload Mass

Total Payload Mass
- 45,596 kg (total) carried by boosters launched by NASA (CRS)

```
 * sqlite:///my_data1.db
Done.
Out[17]:
```

SUM(PAYLOAD_MASS__KG_)
45596

Average Payload Mass
- 2,928 kg (average) carried by booster version F9 v1.1

```
 * sqlite:///my_data1.db
Done.
Out[16]:
```

AVG(PAYLOAD_MASS__KG_)
2928.4

Landing and Mission Info

1st Successful Landing in Ground Pad

- 12/22/2015 •

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

MIN(DATE)
2015-12-22

Booster Drone Ship Landing

Booster mass greater than 4,000

but less than 6,000

- JSCAT-14
- JSCAT-16
- SES-10,
- SES-11 / EchoStar 105

Done.

Payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failed Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

Done.

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

Boosters

Carrying Max Payload

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

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

Failed Landings on Drone Ship

In 2015

- Showing month, date, booster version, launch site and landing outcome

Done.

month	Date	Booster_Version	Launch_Site	Landing_Outcome
01	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Count of Successful Landings

Ranked Descending

- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

Done.

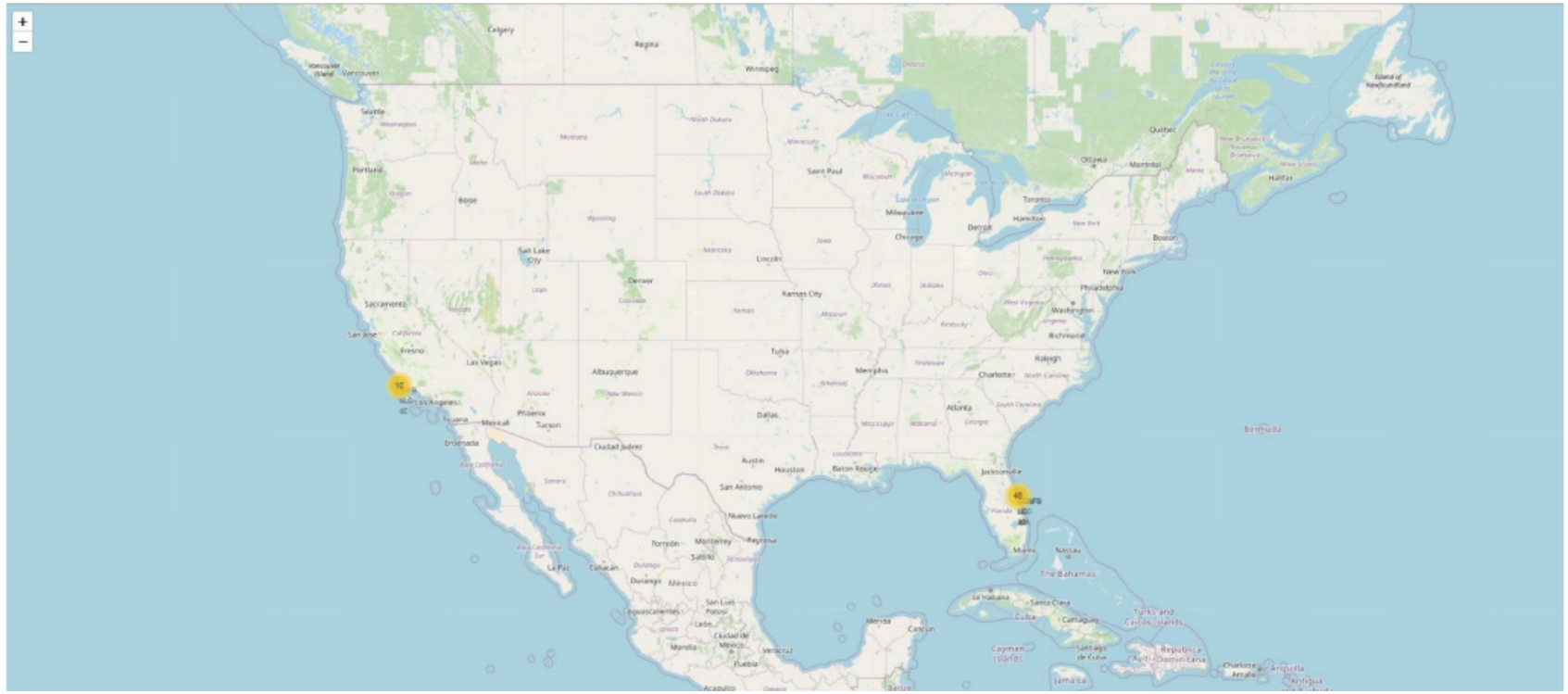
Landing_Outcome	count_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

LAUNCH SITE ANALYSIS

Launch Sites

Near Equator Advantage:

- The closer the launch site to the equator, the easier it is to launch into an equatorial orbit, and the more significant the boost from Earth's rotation for a prograde orbit.
- Rockets launched from sites near the equator receive an additional natural boost due to the rotational speed of Earth. This boost helps save costs by reducing the need for extra fuel and boosters.



Launch Outcomes

At Each Launch Site:

Outcomes:

- Green markers indicate successful launches.
- Red markers indicate unsuccessful launches.

Launch Site Success Rate:

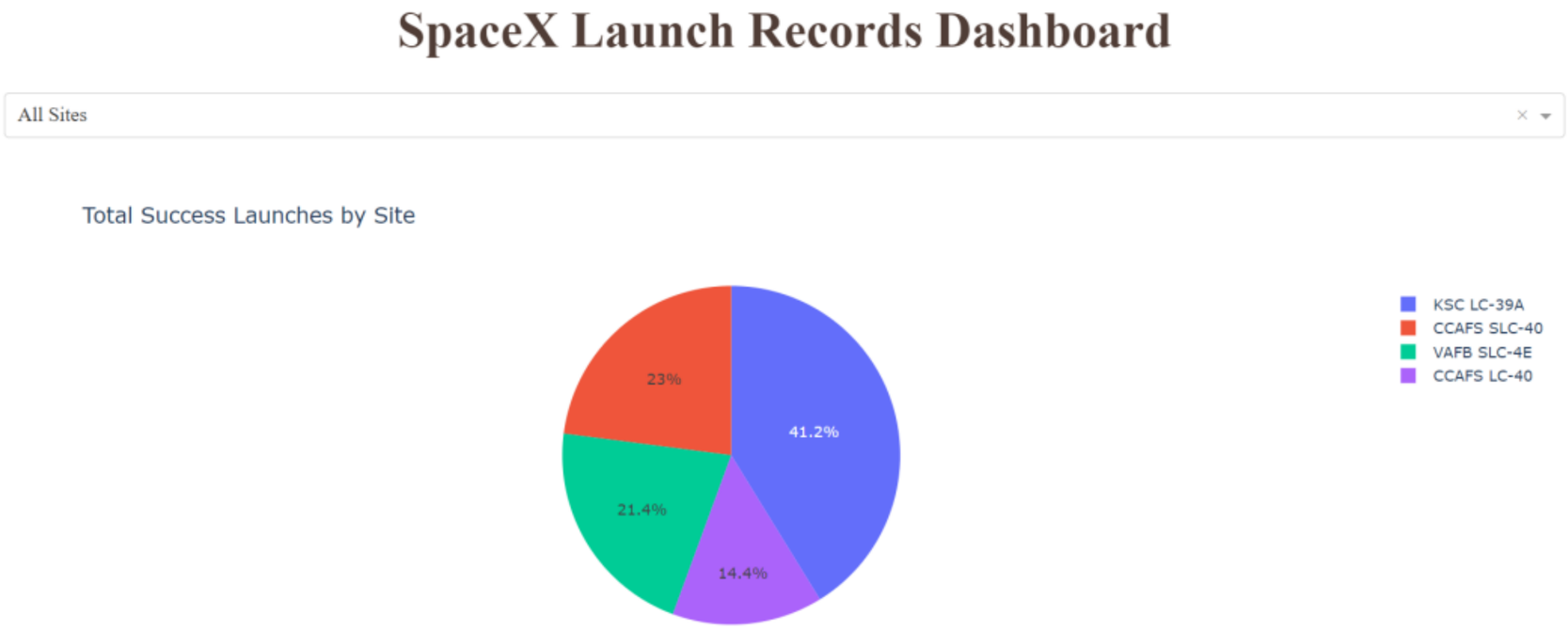
- Launch site CCAFS SLC-40 has a success rate of 3 out of 7 launches (42.9%).

DASHBOARD WITH PLOTLY

Launch Success by Site

Success as Percent of Total

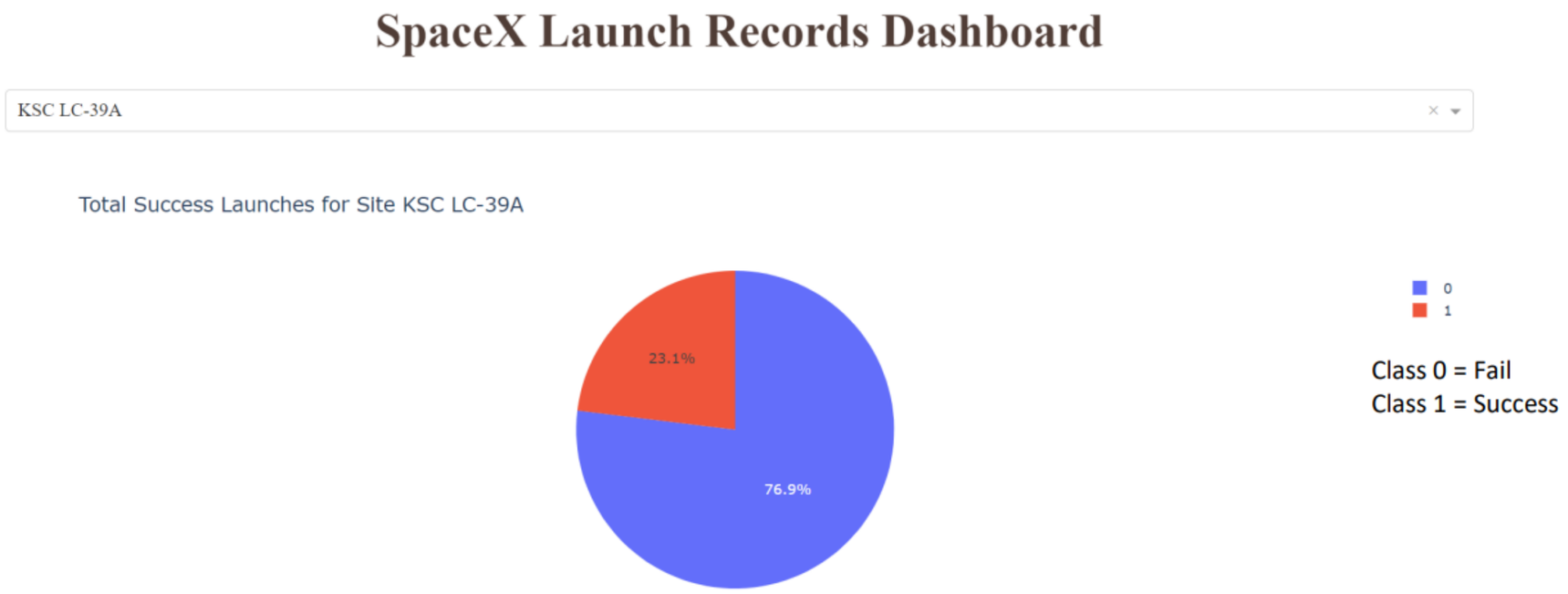
- KSC LC-39A has the most successful launches amongst launch sites (41.2%)



Launch Success (KSC LC-29A)

Success as Percent of Total

- KSC LC-39A has the highest success rate amongst launch sites (76.9%)
- 10 successful launches and 3 failed launches

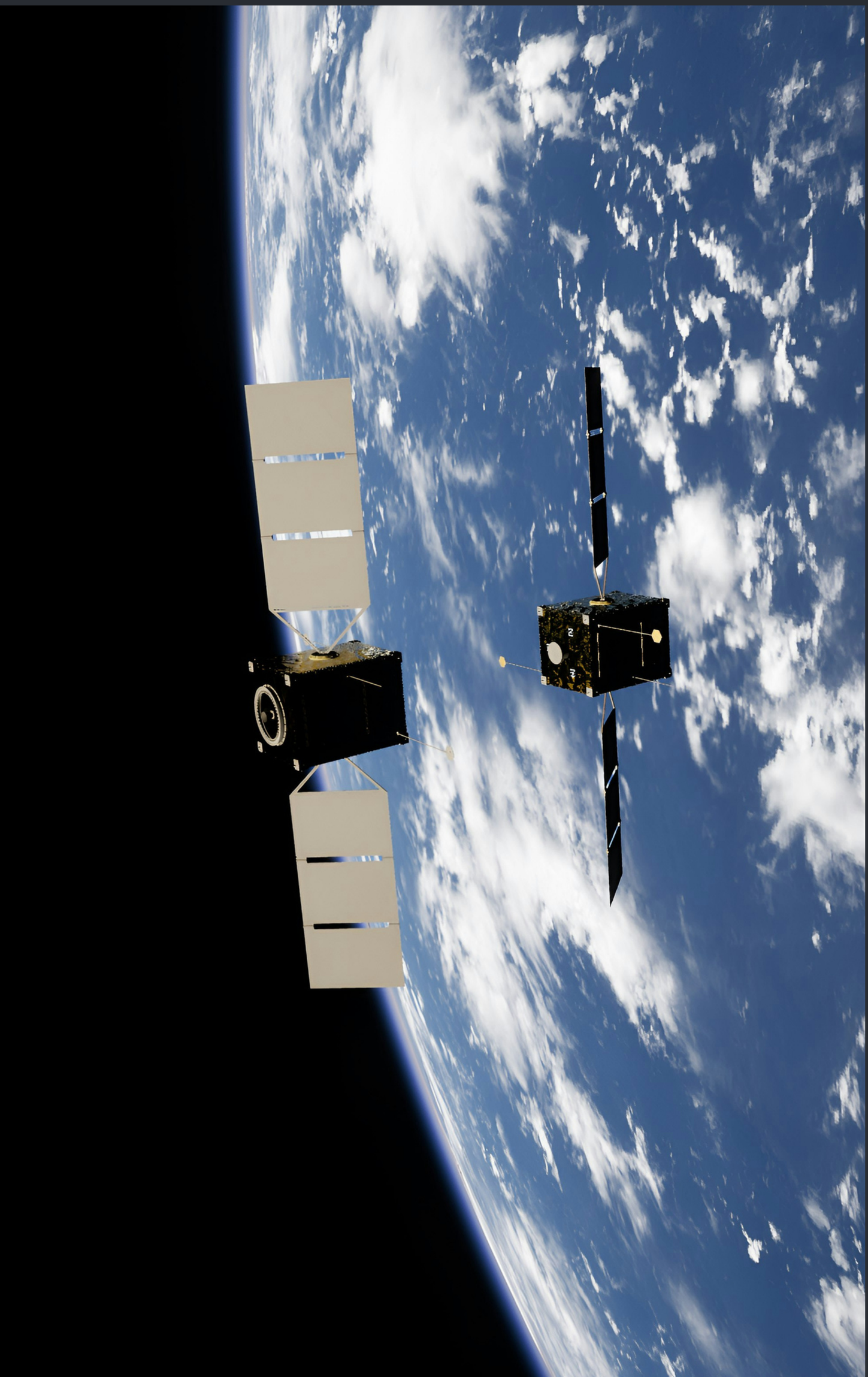


Payload Mass and Success

By Booster Version

- Payloads between 2,000 kg and 5,000 kg have the highest success rate
- 1 indicating successful outcome and 0 indicating an unsuccessful outcome

PREDICTIVE ANALYTICS



Evaluation

Model Performance and Accuracy:

- All models performed similarly and achieved comparable scores and accuracy levels. This consistency is likely attributed to the small dataset, which limits the variability in model performance.
- The Decision Tree model slightly outperformed the other models when considering the `.best_score_` metric, indicating that it achieved a slightly higher average score across all cross-validation folds for a single combination of parameters.

[64]:

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.692308	0.800000
F1_Score	0.888889	0.888889	0.818182	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

[65]:

```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.8589285714285714
Best params is : {'criterion': 'gini', 'max_depth': 5, 'max_features': 'sqrt', 'min_samples_leaf': 1,
'min_samples_split': 5, 'splitter': 'best'}
```

Evaluation with Confusion Matrix

Performance Summary using Confusion Matrix:

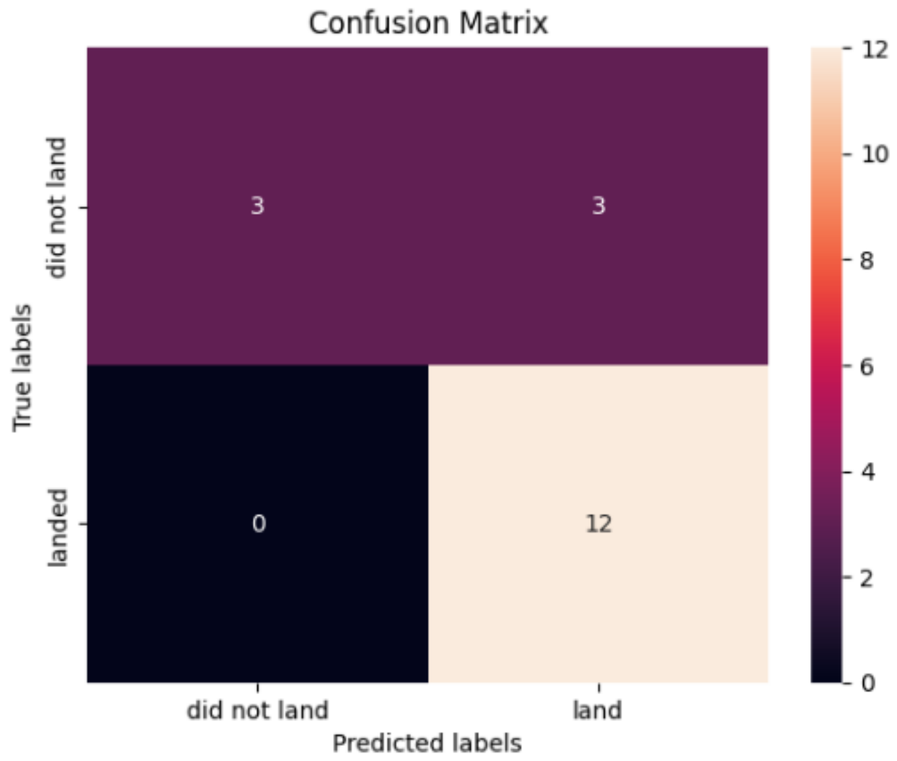
- A confusion matrix provides a detailed summary of the performance of a classification algorithm, showing the counts of true positives, true negatives, false positives, and false negatives.
- In this case, all confusion matrices were identical, indicating consistent model performance across evaluations.

Interpretation of Confusion Matrix Outputs:

- True Positives (TP): 12
- True Negatives (TN): 3
- False Positives (FP): 3
- False Negatives (FN): 0

Metrics Calculations:

- Precision: Precision measures the proportion of correctly predicted positive cases (TP) out of all predicted positive cases (TP + FP). In this case, precision is calculated as 12 / (12 + 3) = 0.80.
- Recall: Recall (also known as sensitivity or true positive rate) measures the proportion of correctly predicted positive cases (TP) out of all actual positive cases (TP + FN). In this case, recall is calculated as 12 / (12 + 0) = 1.00.
- F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. It is calculated as 2 * (Precision * Recall) / (Precision + Recall) = 2 * (0.80 * 1.00) / (0.80 + 1.00) = 0.89.
- Accuracy: Accuracy measures the proportion of correctly predicted cases (both positive and negative) out of all cases. It is calculated as (TP + TN) / (TP + TN + FP + FN) = (12 + 3) / (12 + 3 + 3 + 0) = 0.833, or 83.3%.



Conclusion

Model Performance: The models exhibited comparable performance on the test set, with the decision tree model showing a slight edge in accuracy.

Equator Advantage: Most launch sites are strategically located near the equator to leverage the Earth's rotational speed, which reduces the need for additional fuel and boosters, thus lowering costs.

Coastal Locations: All launch sites are situated near coastal areas, likely for logistical and safety reasons, facilitating easier transportation of equipment and providing a safe area for launches.

Launch Success Trend: The success rate of launches has shown a positive trend over time, indicating improvements in technology, processes, and operational efficiency.

KSC LC-39A Performance: KSC LC-39A stands out with the highest success rate among launch sites. Notably, it achieves a perfect 100% success rate for launches with payloads under 5,500 kg, highlighting its reliability for smaller missions.

Orbital Success Rates: Specific orbits, such as ES-L1, GEO, HEO, and SSO, have achieved a 100% success rate, indicating a high level of precision and reliability in placing satellites into these orbits.

Payload Mass Impact: Across all launch sites, there is a clear correlation between payload mass and launch success rate, with higher payload masses associated with higher success rates. This suggests that launch vehicles are well-designed to handle heavier payloads more reliably.

Conclusion

Dataset Expansion: Expanding the dataset to include more observations could enhance the robustness of the predictive analytics results. This expansion would enable researchers to assess the generalizability of the findings to a broader range of scenarios and conditions, potentially increasing the reliability of the conclusions drawn from the study.

Feature Analysis and PCA: Conducting further analysis on the features or applying principal component analysis (PCA) could offer insights into improving the accuracy of the models. Feature analysis helps identify the most relevant variables for prediction, while PCA can reduce the dimensionality of the dataset, potentially improving model performance and interpretability.

Exploration of XGBoost: The study did not explore the use of XGBoost, a powerful and popular gradient boosting algorithm. Including XGBoost in the analysis would be valuable to determine if it outperforms the other classification models considered in the study. XGBoost's ability to handle complex relationships and large datasets could lead to improved model performance and predictive accuracy.