Head-to-head with Space X

BARIS SAYGILI

29-06-2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion
- Appendix

EXECUTIVE SUMMARY

1. The author collected data from the public SpaceX API and SpaceX Wikipedia page, creating a labeled column to classify successful landings.
2. Data exploration and visualization techniques were used, including SQL queries, visualizations, folium maps, and dashboards.
3. Relevant columns were selected as features, categorical variables were converted to binary using one-hot encoding, and the data was standardized.

4.Four machine learning models (Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors) were built and evaluated, showing similar accuracy rates of about 83.33% but over-predicting successful landings, indicating a need for more data for improved accuracy.

INTRODUCTION

- **Situation:**

•Space X offers the best pricing in the industry, with a cost of $62 million compared to $165 million USD for other providers.

- One of the key reasons for Space X's competitive pricing is its ability to recover and reuse the first stage of its rockets.

- **Problem:**

- Space Y wants us to develop a machine learning model for predicting the success of Stage 1 rocket recovery.

METHODOLOGY

- Data collection:

- Data was gathered from both SpaceX's public API and their Wikipedia page.

- Data preprocessing:

- Data was cleaned and transformed through the process of data wrangling.

- Labeling:

- A classification was performed to categorize landings as either successful or unsuccessful.

- Exploratory data analysis:

- Exploratory data analysis was conducted using SQL queries and visualization techniques.

- Interactive visual analytics:

- Folium and Plotly Dash were utilized for interactive visual analytics.

- Predictive analysis:

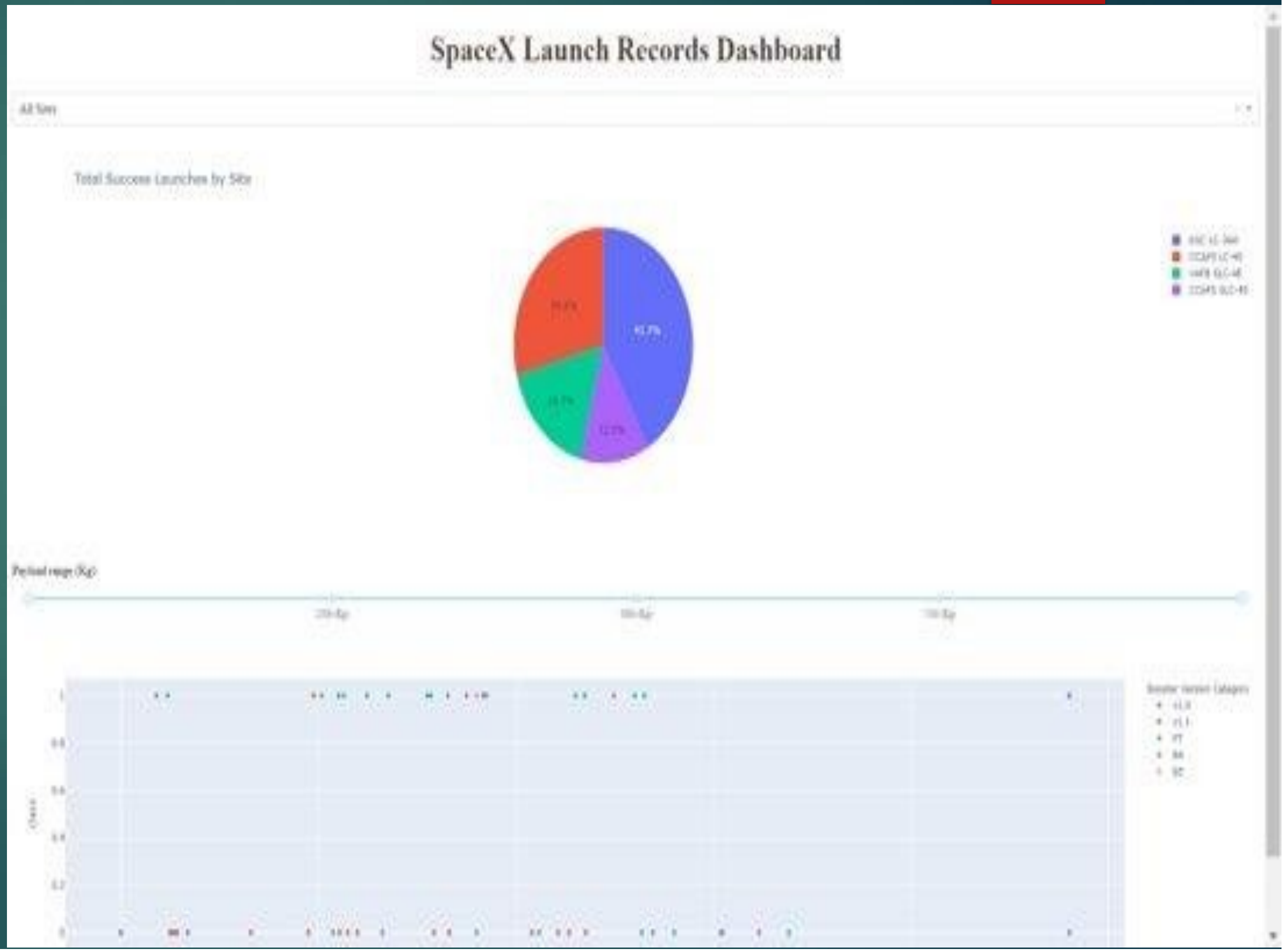
- Classification models were employed for predictive analysis.

- Model tuning:

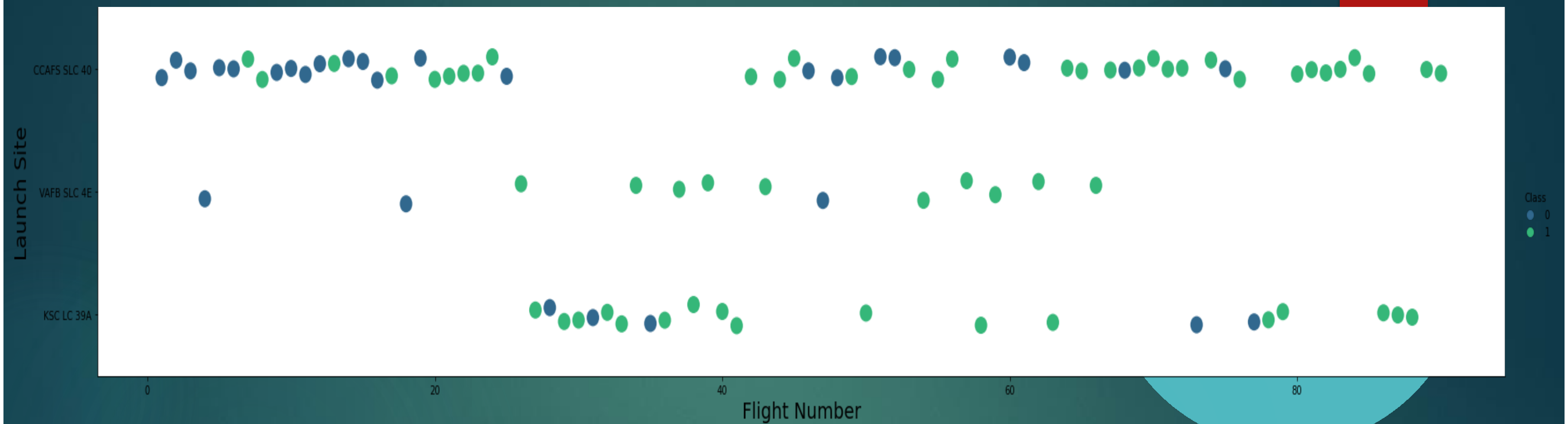
- GridSearchCV was utilized to fine-tune the models.

RESULTS

The Plotly dashboard provides a preview of the following sections: EDA visualization, EDA with SQL, Interactive Folium map, and the model results with an accuracy of approximately 83%.



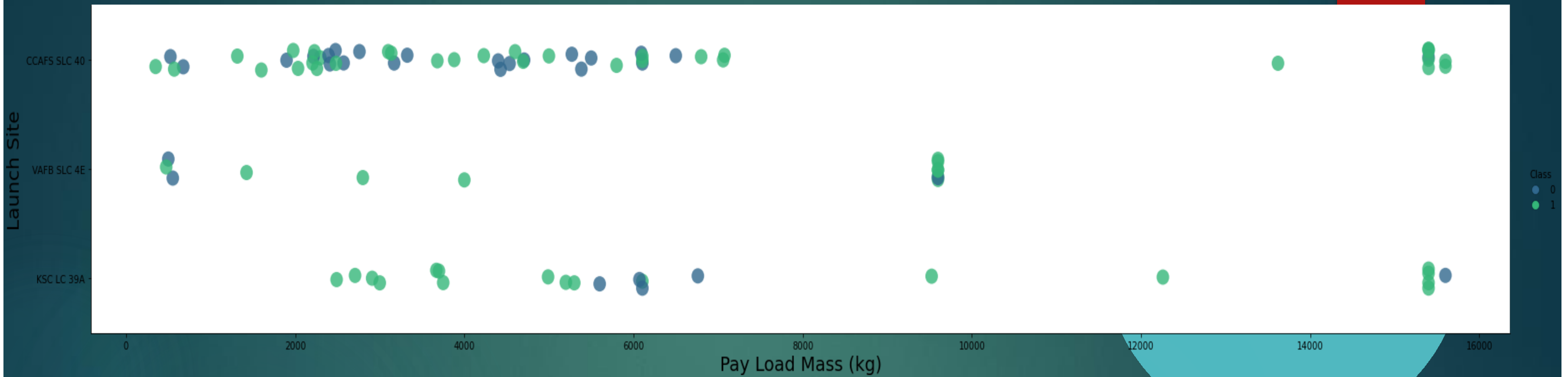
Flight Number vs. Launch Site



Green : Successful launch
Dark Blue: Unsuccessful launch

The graphic indicates a progressive increase in the success rate over time, particularly around flight number 20, which marked a significant breakthrough. CCAFS emerges as the primary launch site with the highest volume.

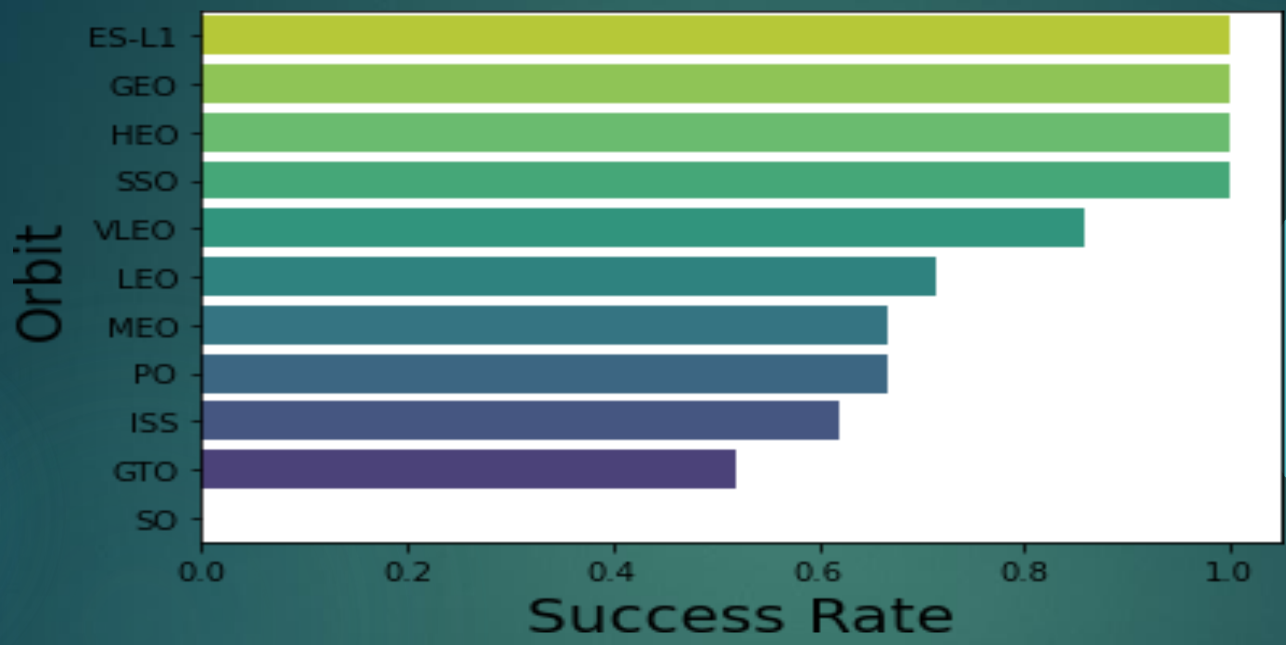
Payload vs. Launch Site



Green : Successful launch
Dark Blue: Unsuccessful launch

Payload mass is predominantly distributed within the range of 0-6000 kg, and different launch sites exhibit variations in the payload mass they handle.

Success rate vs. Orbit type



Success Rate Scale with
0 as 0%
0.6 as 60%
1 as 100%

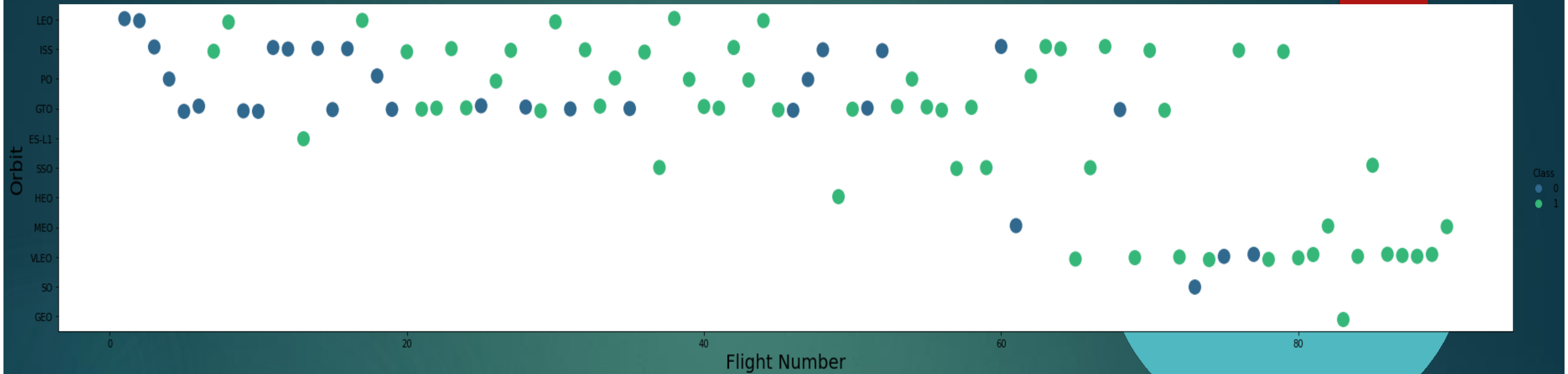
ES-L1 , GEO , HEO ,SSO have 100% success rate

VLEO has decent success rate and attempts

SO has 0% success rate

GTO has the around 50% success rate but largest sample

Flight Number vs. Orbit type

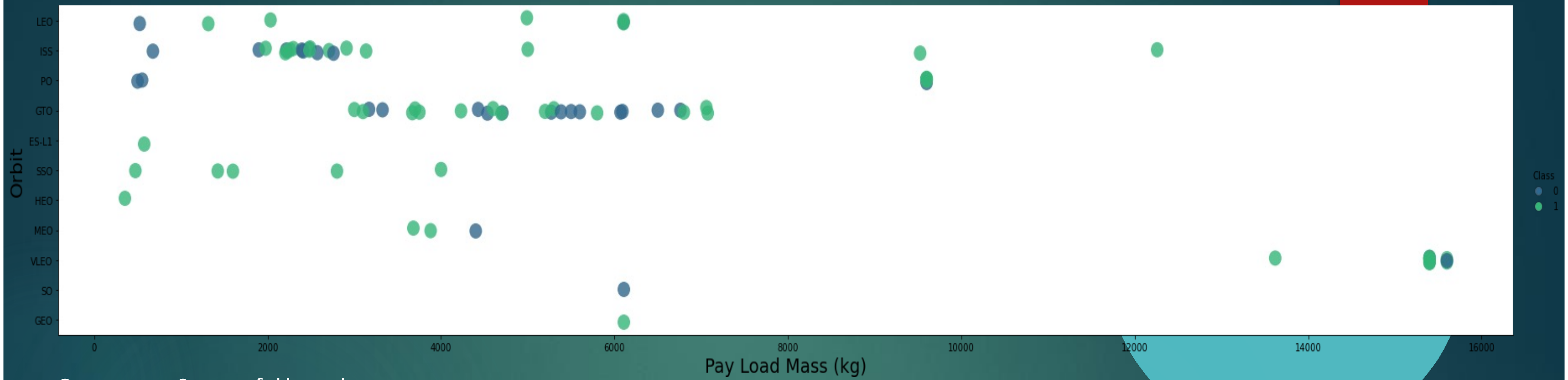


Green : Successful launch

Dark Blue: Unsuccessful launch

- Launch orbit preferences have changed over flight numbers.
- There is a correlation between launch outcome and these preferences.
- Initially, SpaceX focused on LEO orbits, which had moderate success.
- In recent launches, there has been a shift back to VLEO.
- SpaceX performs better in lower orbits and Sun-synchronous orbits.

Payload vs. Orbit type

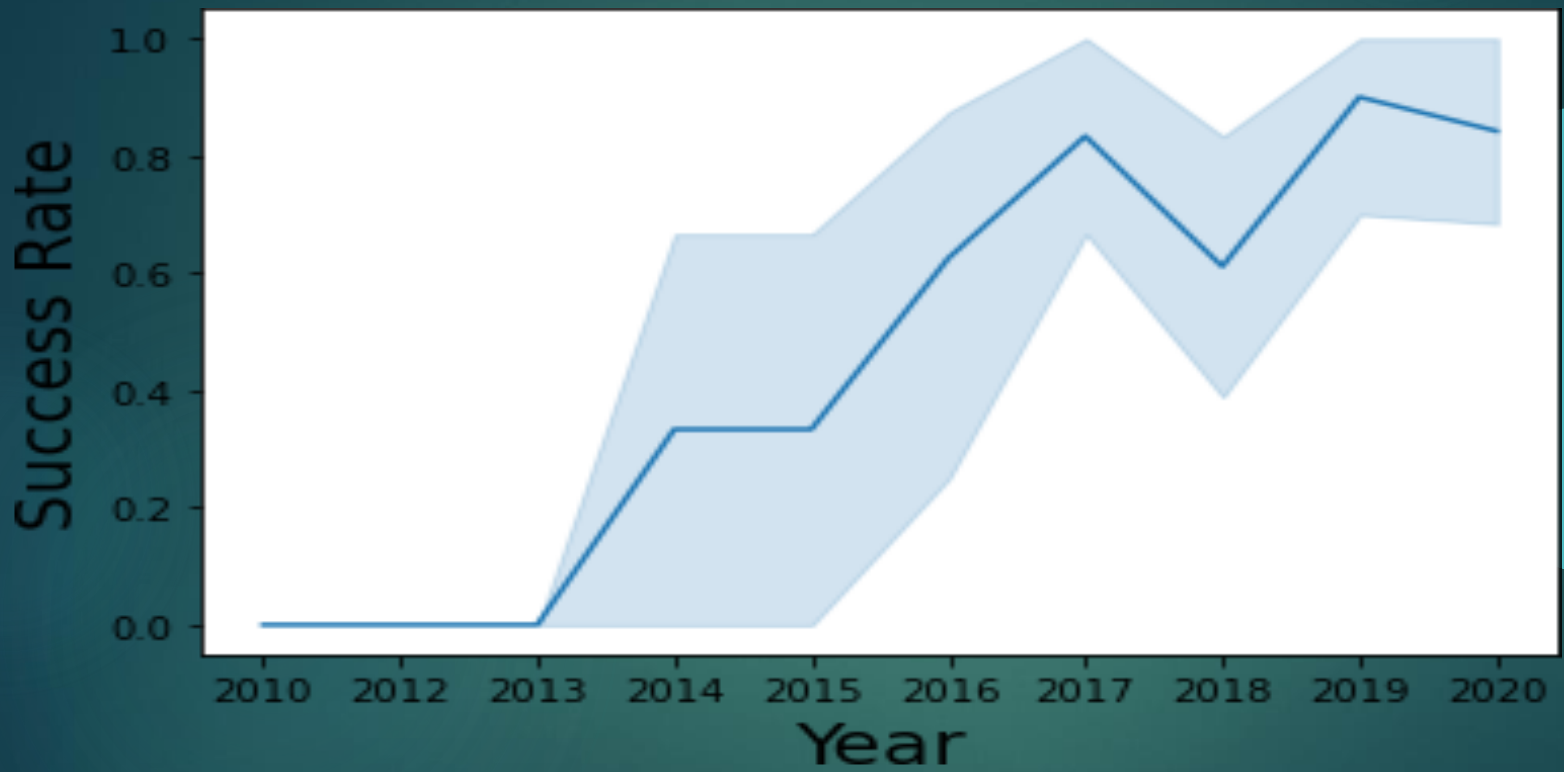


Green : Successful launch

Dark Blue: Unsuccessful launch

There is a correlation between payload mass and orbit choice. LEO and SSO orbits tend to have lower payload masses, while the VLEO orbit, which is the one of the most successful, typically has higher payload mass values.

Launch Success Yearly Trend



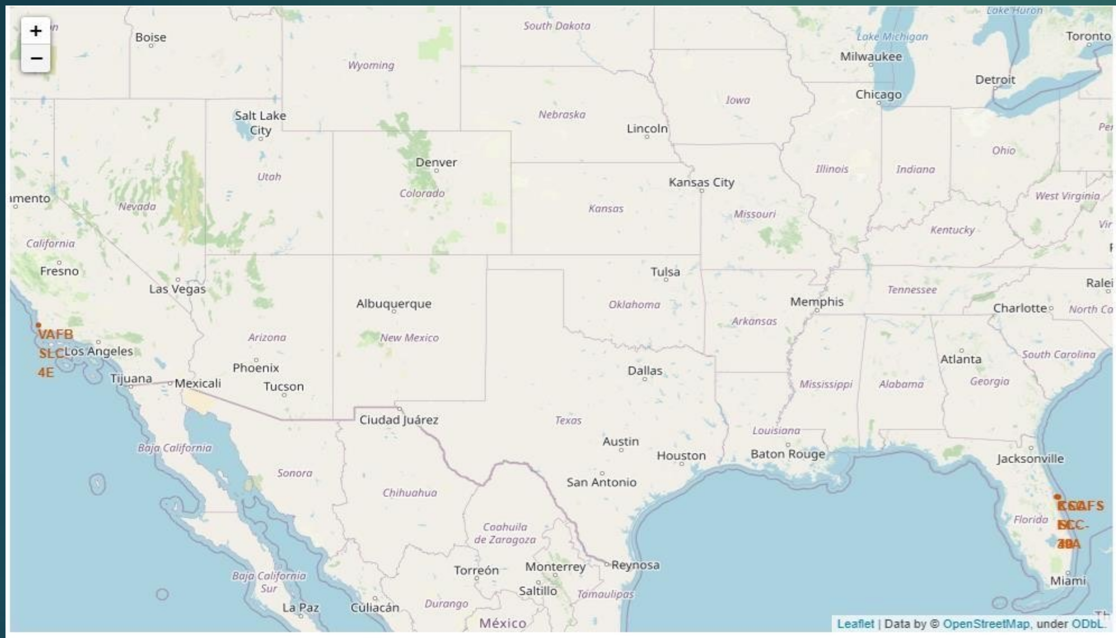
95% confidence interval (light blue shading)

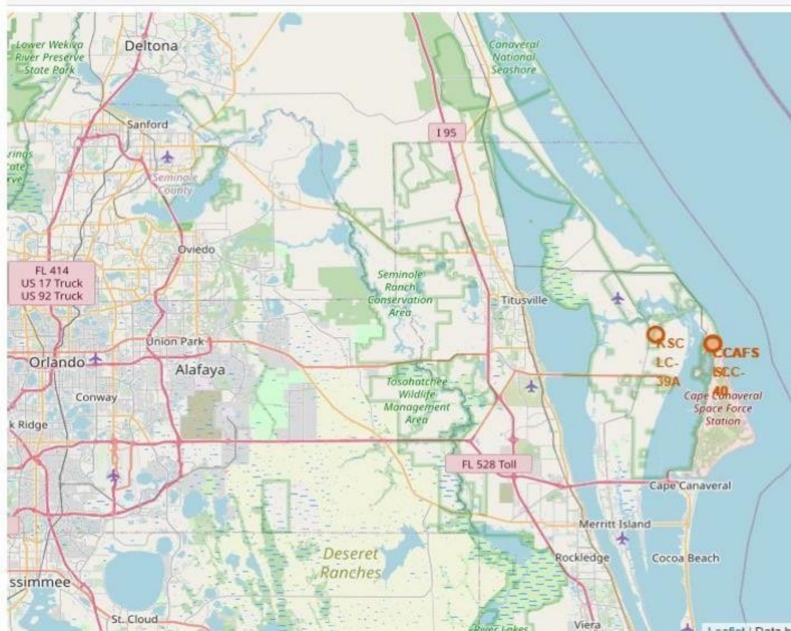
Success generally increases over time since 2013

Success in recent years at around 80%

Let's Check Out the Locations

Launch Site Locations





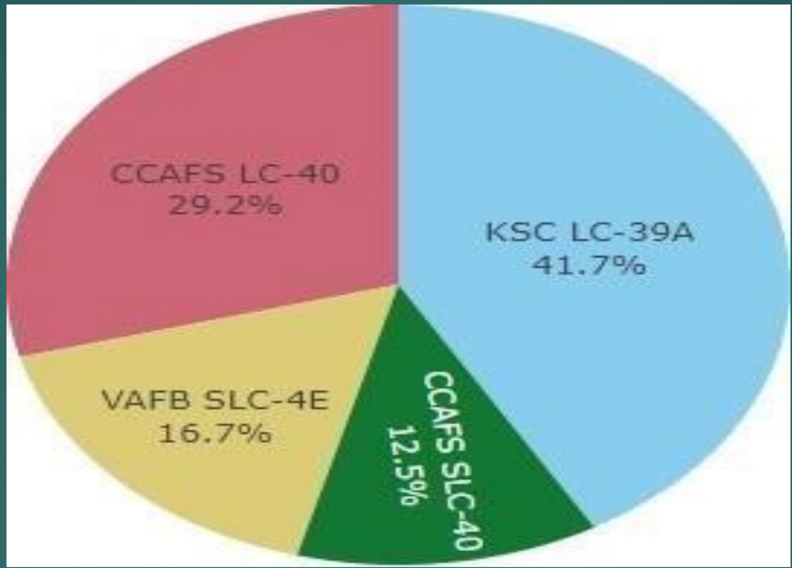
The left map displays all launch sites on a US map, while the right map focuses specifically on the two launch sites in Florida due to their proximity. It is worth noting that all launch sites are situated near the ocean.

Color-Coded Launch Markers

The clusters on the Folium map can be interactively clicked to reveal successful landings (green icons) and failed landings (red icons). For instance, the VAFB SLC - 4E launch site in this example displays 4 successful landings and 6 failed landings.

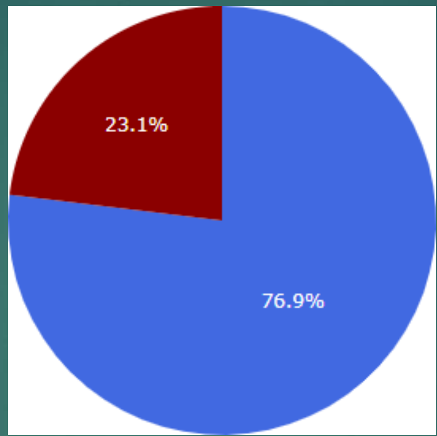
Dashboards

Successful Launches Across Launch Sites



The distribution of successful landings across different launch sites reveals some interesting patterns. CCAFS LC-40, which is now known as CCAFS SLC-40, and KSC have an equal number of successful landings. However, the majority of these successful landings occurred before the name change. In contrast, VAFB has the smallest share of successful landings, which could be attributed to a smaller sample size and the increased challenges associated with launching from the west coast.

Highest Success Rate Launch Site



KSC LC-39A Success Rate (blue=success)

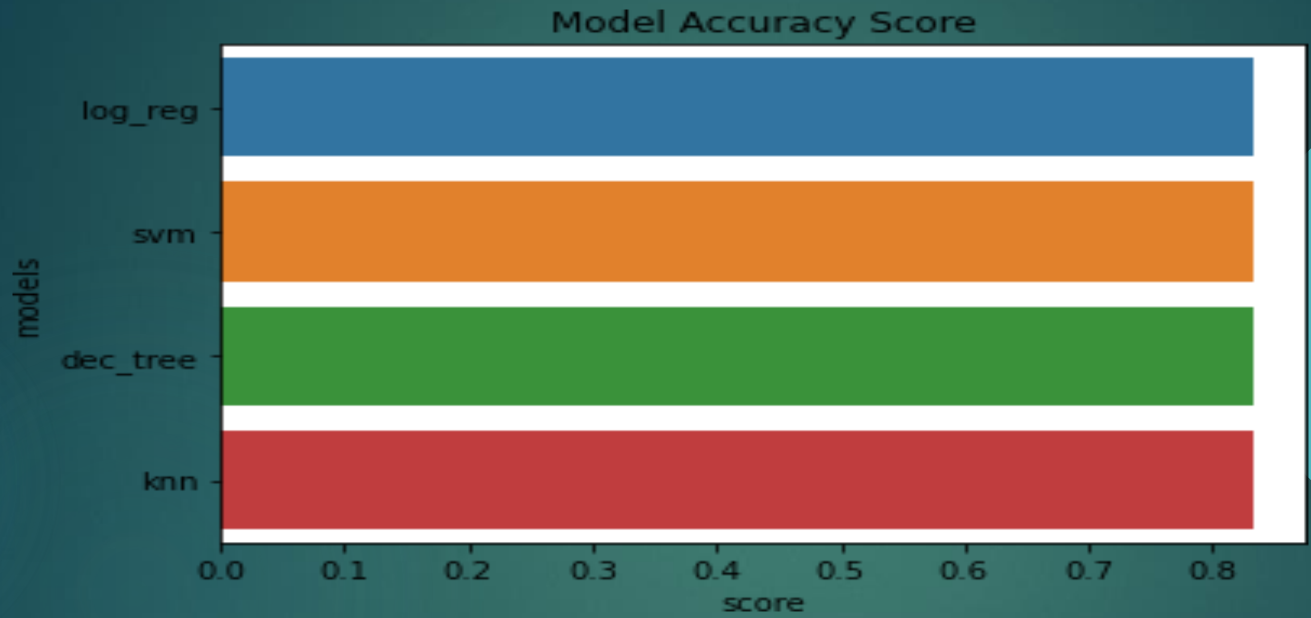
KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings

Payload Mass vs. Succes vs. Booster Version Category

The Plotly dashboard includes a payload range selector, but it is currently set from 0 to 10000 instead of the maximum payload of 15600. The scatter plot incorporates the booster version category as color and the number of launches as point size. Within the payload range of 0-6000, it is noteworthy that two failed landings had payloads of zero kg.

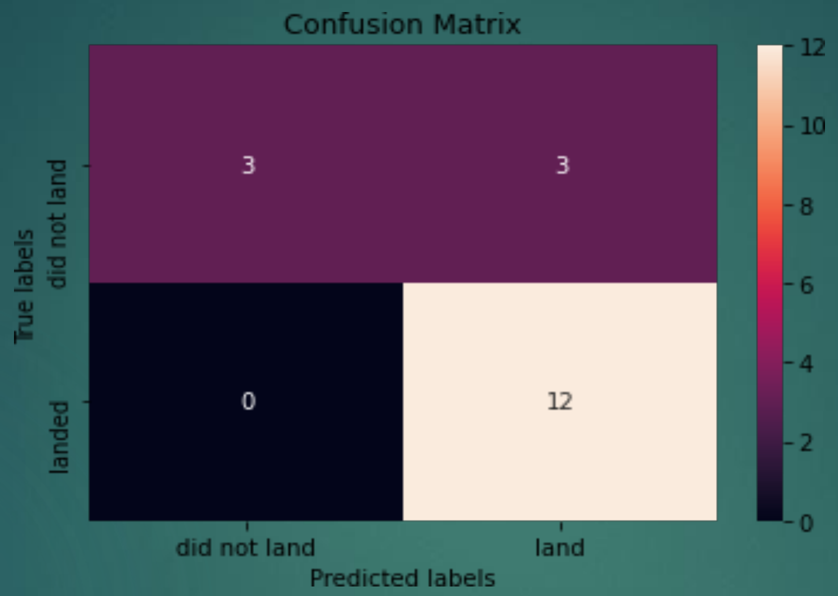
Predictive Analysis (Classification)

Classification Accuracy



All models achieved a similar accuracy of 83.33% on the small test set.
The small test size, consisting of only 18 samples, can lead to significant variance in accuracy results.
The Decision Tree Classifier model specifically showed inconsistent accuracy in repeated runs.
To determine the best model and improve accuracy, gathering more data is likely necessary.

Confusion Matrix



The correct predictions form a diagonal pattern from the top left to the bottom right.

The models correctly predicted 12 successful landings when the true label was a successful landing.

The models correctly predicted 3 unsuccessful landings when the true label was an unsuccessful landing.

The models incorrectly predicted 3 successful landings when the true label was an unsuccessful landing, resulting in false positives.

Overall, the models tend to overpredict successful landings, indicating a bias towards positive predictions.

Conclusion

- - Our task was to develop a machine learning model for Space Y, a competitor bidding against SpaceX.
- - The objective of the model was to predict the successful landing of Stage 1, resulting in savings of approximately $100 million USD.
- - To achieve this, we utilized data from a public SpaceX API and performed web scraping on the SpaceX Wikipedia page.
- - Data labeling was conducted, and the collected data was stored in a DB2 SQL database.
- - A dashboard was created for visualization purposes.
- - The machine learning model we developed achieved an accuracy of 83%.
- - This model can be utilized by Elon Musk and Space Y to predict, with a relatively high accuracy, whether a launch will have a successful Stage 1 landing before the actual launch, enabling them to make informed decisions regarding the launch.
- - Collecting more data would be beneficial to further determine the optimal machine learning model and enhance accuracy.

Appendix

- GitHub repository url:
- https://github.com/barissaygili/Capstone_Project

- Presenter:
- BARIS SAYGILI