The Future of Space Exploration: Analyzing SpaceX's Rocket Reusability

Chahboune Ahmed

06/17/2023

OUTLINE

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

EXECUTIVE SUMMARY

- **Methodology:**

- Data Collection through API.
- Data Collection wits Web Scraping.
- Data Wrangling.
- Exploratory Data Analysis wits SQL.
- Exploratory Data Analysis wits Data Visualization.
- Interactive Visual Analytics wits Folium.
- Machine Learning Prediction

- **Results:**

- Exploratory Data Analysis result.
- Interactive Analytics result.
- Predictive Analytics result.

INTRODUCTION

In this project, web will explore different methodologies for data analysis and prediction. Our main focus will be on collecting data through APIs and web scraping, cleaning and preparing the data for analysis, and conducting explore tory data analysis using SQL and data visualization techniques. We will also delve into interactive visual analytics using Folium for geographical analysis and apply machine learning algorithms for predictive analytics. The goal is to extract insights, identify patterns, and make predictions based on the data. At the end, web will summarize all our findings to provide a clear and concise summary of the results obtained throughout the project.

METHODOLOGY

- **Summary**:
- Data was collected using SpaceX API and web scraping from Wikipedia.
- Data wrangling techniques were applied, including one-hot encoding for categorical features.
- Exploratory data analysis (EDA) was conducted using visualization and SQL.
- Interactive visual analytics were performed using Folium and Plotly Dash.
- Predictive analysis focused on classification models.
- The process of building, tuning, and evaluating the classification models is explained.

Data Collection Using API/Web Scarping

Using SpaceX API:

Start requesting rocket launch data from SpaceX API using the requests module.

response = requests.get(spacex_url)

Decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize() method.

Using Web Scraping:

Extract a Falcon 9 launch record HTML table from Wikipedia wits BeatifulSoup.

soup = BeautifulSoup(response.content, 'html.parser')

Parse the table and convert it into a Pandas data frame.

html_tables = soup.find_all('table')

The link to the notebook:

API: https://github.com/med04/Applied-Data-Science-Capstone/blob/main/spacex-data-collection-api.ipynb

Web Scraping: https://github.com/med04/Applied-Data-Science-Capstone/blob/main/spacex-data-collection-webscraping.ipynb

Data Wrangling

Refers to the process of cleaning, transforming, and preparing the collected data for analysis.

Check the missing values of rows in the dataset web use the .isnull() method.

Calculate the number of launches on each site.

Calculate the number and occurrence of each orbit.

Calculate the number and occurrence of mission outcome per orbit type.

Create a landing outcome label from Outcome column.

Exploratory Data Analysis(EDA) wits SQL

Wrote and execute SQL queries to get insight the data including:

- The names of the unique launch sites in the space mission.
- The total payload mass carried by boosters launched by NASA (CRS).
- The average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved.
- List the names of the booster which have success in drone ship and have a payload mass greater than 4000 but les than 6000.
- The total number of successful and failure mission outcomes.
- The names of the booster version which have carried the maximum payload mass.
- Count of successful landing outcomes between the date 04-06-2010 and 20-03-2017.

The link to the notebook: https://github.com/med04/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-sql-

EDA wits Visualization

With exploratory Data Analysis with visualization Tools like matplotlib and seaborn, we can determine:

- How the Flight Number and Payload variables would affect the launch Outcome.
- Which Orbit have high success rate.
- if there is any relationship between Flight Number and Orbit type.
- When the success rate starts increasing.
- How each important variable would affect the success rate.

This analysis will help us select the features that will be used in success prediction in the future.

After the selction, some variables are categorical and should be converted to numerical values. To achieve this we apply OneHotEncoder using the function get_dummies() on the categorical columns.

The Link to the notebook:

https://github.com/med04/Applied-Data-Science-Capstone/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

Interactive Visual Analysis with Folium

Folium can help us interact with maps in a way to:

- Mark all launch sites on a map.
- Add highlighted circle area for each launch site
- Create markers for all launch sites that was successful or not.
- Create marker cluster having the same coordinate.
- Add a MousePosition on the map to get coordinate for a mouse over a point on the map.
- Calculate the distance between the launch site and the coastline.
- Draw a PolyLine between a Launch site to the selected coastline point.

The Link to the notebook:

https://github.com/med04/Applied-Data-Science-Capstone/blob/main/lab_jupyter_launch_site_location.jupyterlite.ipynb

Interactive Dashboard

Dashboards are a great way to obtain some insights to answer questions, we will use it to determine :

- Which site has the largest successful launches.
- Which payload range has the highest launch success rate.
- Which payload range has the lowest launch rate
- Which Falcon 9 Booster version has the highest launch success rate.

The Link to the Python code: https://github.com/med04/Applied-Data-Science-Capstone/blob/main/dash_app.py

Predictive Analysis

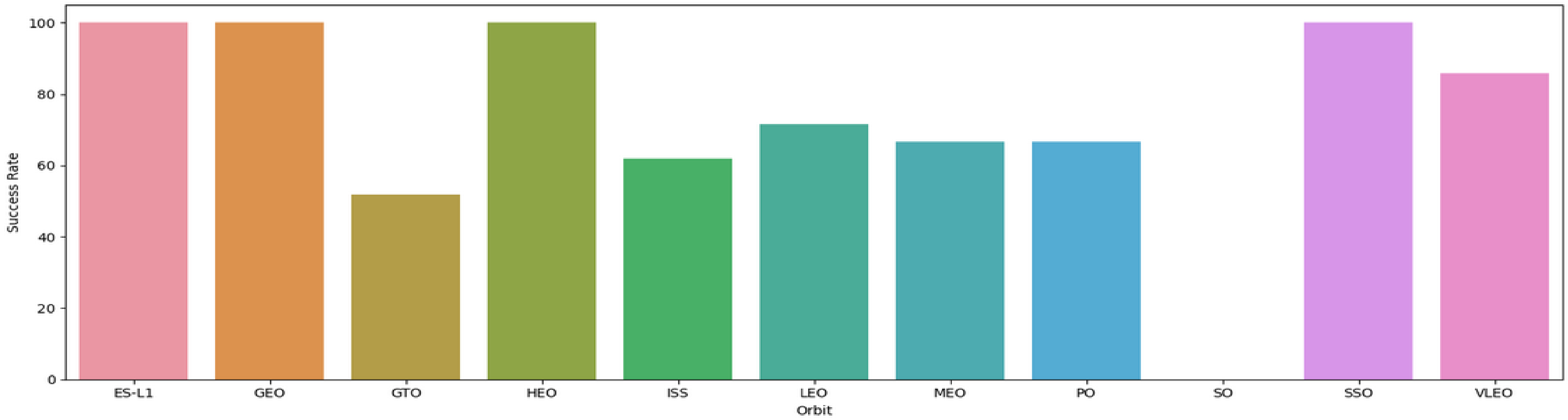
The aim of predictive analysis is to perform exploratory data analysis to determine Training Labels.

Find best Hyperparameters to find the method that performs best using test data.

Use metrics to evaluate different methods like Grid Search Cross Validation, Confusion matrix, etc.

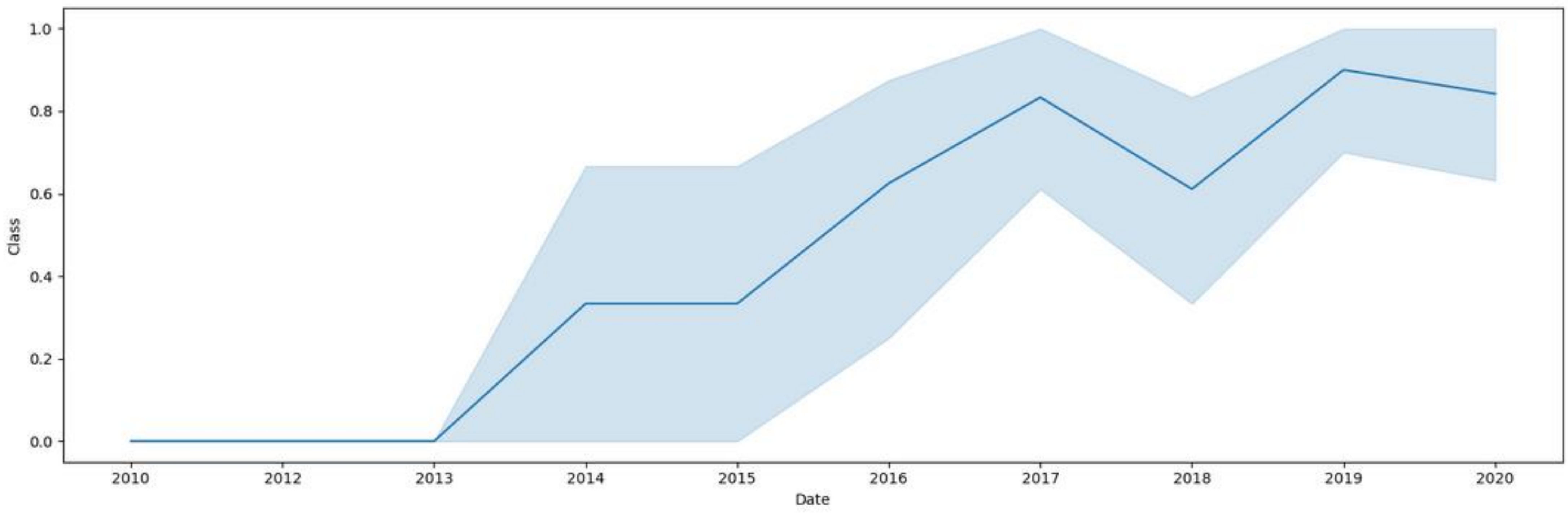
The Link to the notebook: https://github.com/med04/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Exploratory Data Analysis result



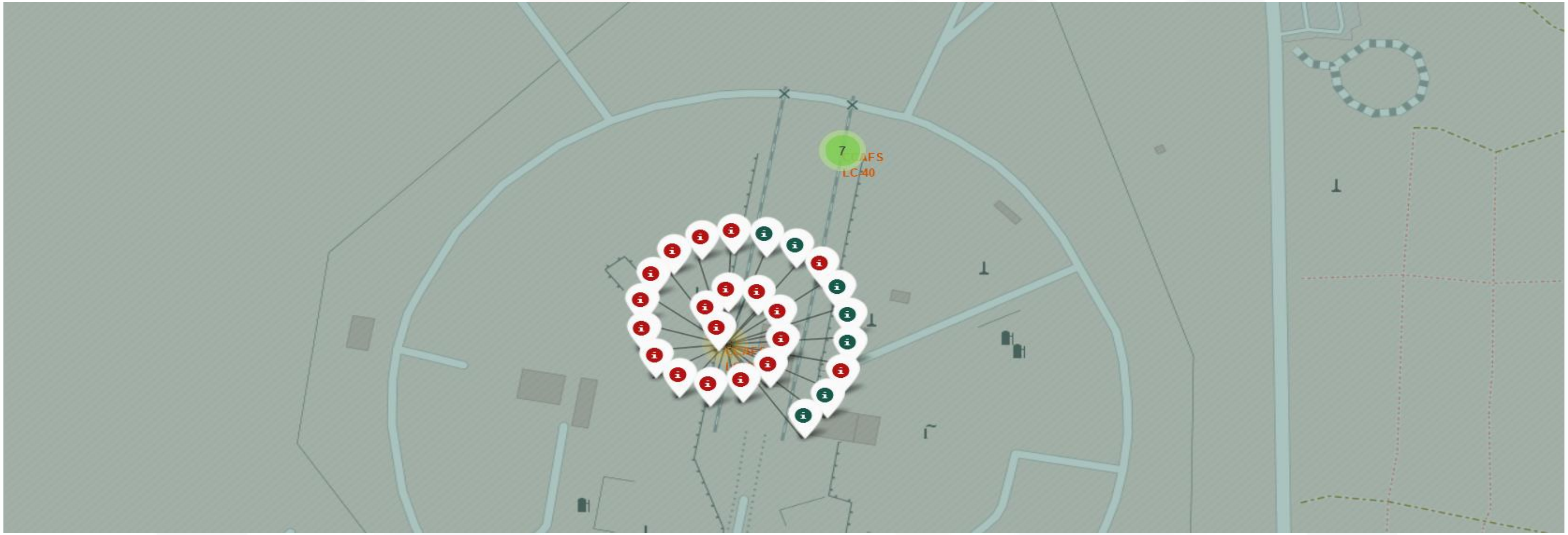
We can see the ES-L1, GEO, HEO and SSO the orbits that have the highest success rate.

Exploratory Data Analysis result



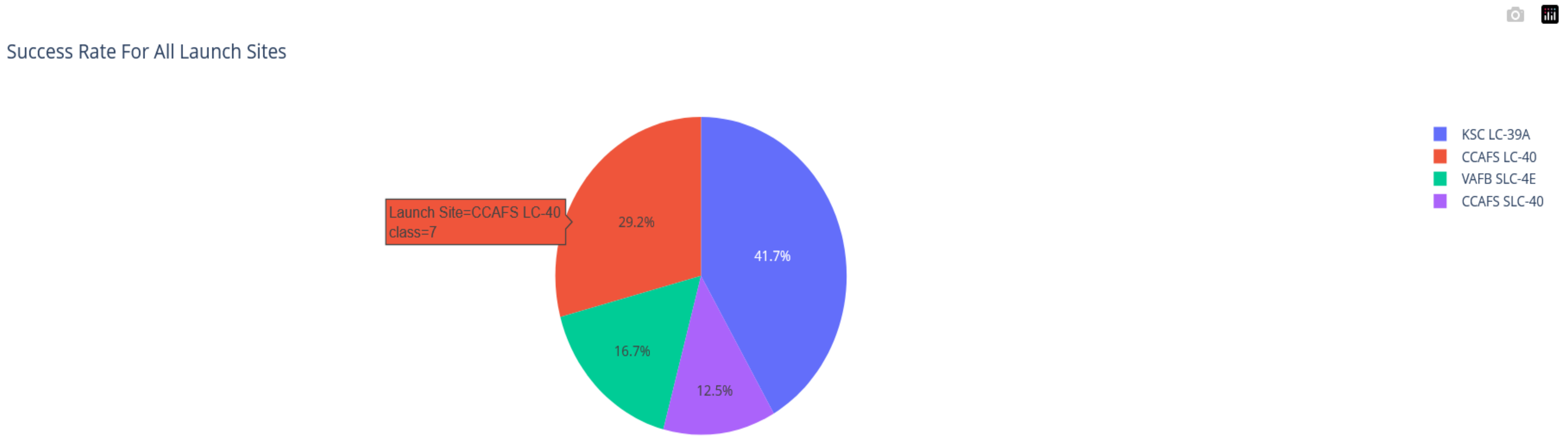
We can observe that the success rate since 2013 kept increasing till 2020

Interactive Visual Analytics result



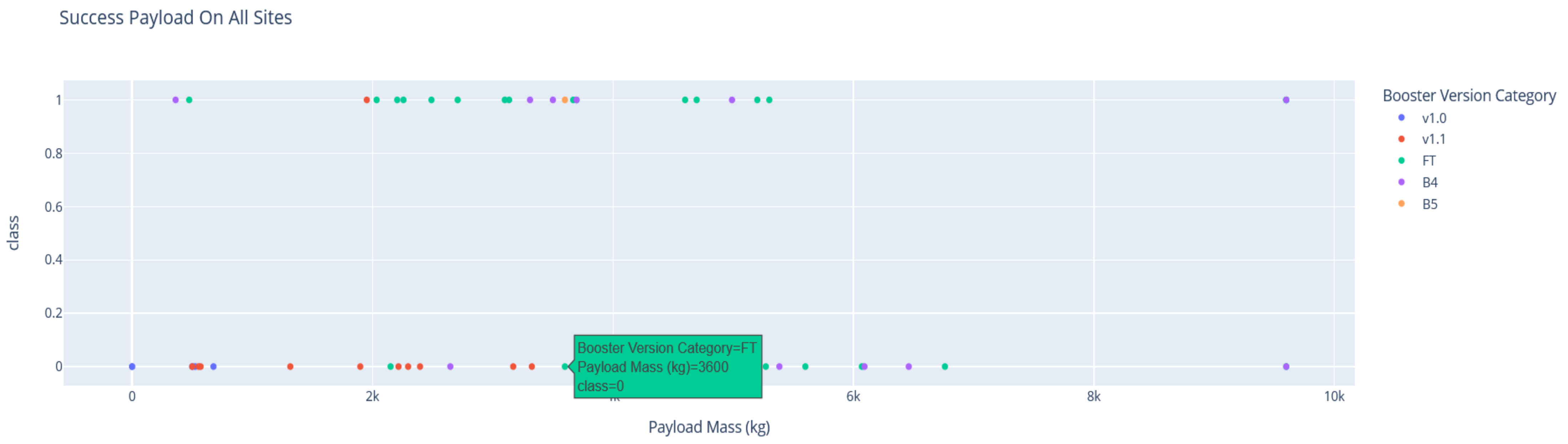
We can also see the success rate of each launch site using Folium, for example the site figured here is for CCAFS LC-40 which has more failure launches than the success one.

Interactive Analytics result



We see that the largest successful launches is KSCC LC-39A for all sites.

Interactive Analytics result



From the scatter plot which show the relationship between the Payload Mass and The class for different Booster Version Categories, we can see that the Payload Mass between 2000kg and 6000 Kg has the highest Success rate (FT Booster Version).

Predictive Analytics result

```
[53]: best_score_models = {'Logistic Regression' : logreg_cv.best_score_,
                         'Support Vector Machine' : svm_cv.best_score_,
                         'Decision Tree' : tree_cv.best_score_,
                         'K-Nearest Neighbors' : knn_cv.best_score_}

      best_model = max(best_score_models, key=best_score_models.get)
      print(f'Best model is {best_model} with a score of {best_score_models[best_model]}')

      Best model is Decision Tree with a score of 0.8785714285714284
```

After the evaluation of all models, as a result we conclude that Decision Tree model is the best to perform well with a score of 0.87.
The Link of the notebook: https://github.com/med04/Applied-Data-Science-Capstone/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

CONCLUSION

- In summary, this project successfully utilized various methodologies, including data collection through API and web scraping, data wrangling, exploratory data analysis (EDA) with visualization and SQL, interactive visual analytics with Folium and Plotly Dash, and predictive analysis using classification models. These methodologies provided valuable insights, patterns, and predictions based on the data, contributing to a deeper