



Space Y First stage reuse

Matheus de Oliveira

7/31/2023

Contents

Executive Summary

Introduction

Methodology

Results

- EDA With visualization

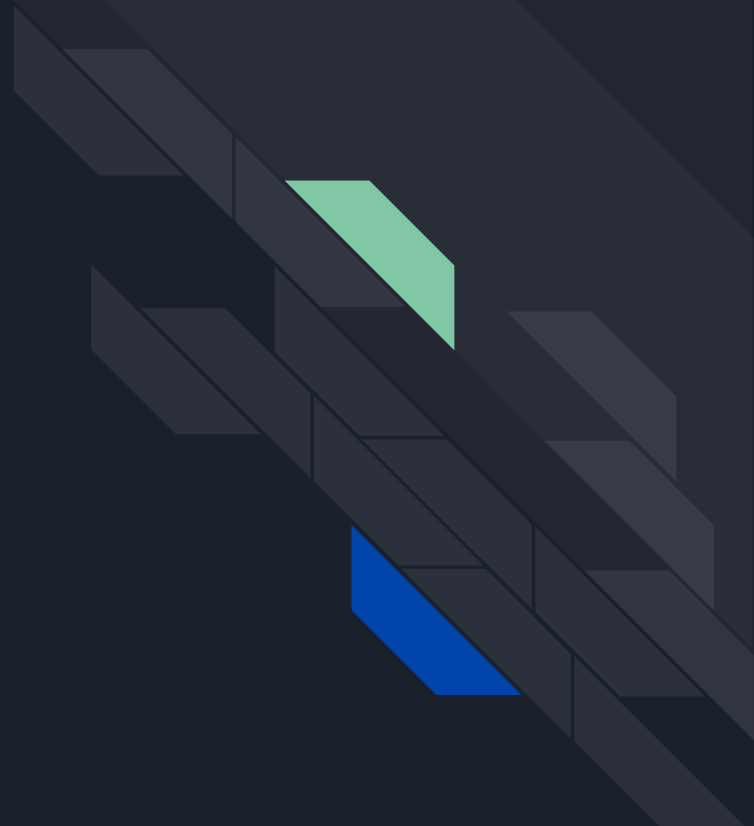
- EDA with SQL

- Interactive maps with folium

- Plotly dashboard

- Predictive analysis

Conclusion





Executive summary

The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies were used:

- **Collect** data using SpaceX REST API and web scraping techniques
- **Wrangle** data to create success/fail outcome variable
- **Explore** data with data visualization techniques, considering the following factors: payload, launch site, flight number and yearly trend
- **Analyze** the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- **Explore** launch site success rates and proximity to geographical markers
- **Visualize** the launch sites with the most success and successful payload ranges
- **Build Models** to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)

The results are:

- **Exploratory Data Analysis:**
 - Launch success has improved over time
 - KSC LC-39A has the highest success rate among landing sites
 - Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- **Visualization / Analytics:**
 - Most launch sites are near the equator, and all are close to the coast
- **Predictive Analytics:**
 - All models performed similarly on the test set. The decision tree model slightly outperformed when looking at `.best_score_`



Introduction

Background

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive (\$62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of \$165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

We will **explore**:

- 01 How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- 02 Rate of successful landings over time
- 03 Best predictive model for successful landing (binary classification)



Methodology

The steps are:

- 1) Collect
- 2) Wrangle
- 3) Explore
- 4) Visualize
- 5) Build Models






Methodology

Data collection - API

The steps responsible to get data from API are:


- Request data from SpaceX API (rocket launch data)
 - Decode response using `.json()` and convert to a dataframe using `.json_normalize()`
 - Request information about the launches from SpaceX API using custom functions
 - Create dictionary from the data]
 - Create dataframe from the dictionary
 - Filter dataframe to contain only Falcon 9 launches
 - Replace missing values of Payload Mass with calculated `.mean()`
 - Export data to csv file
- 



Methodology

Data collection - Web scrapping

The steps responsible to get data from web scrapping are:

- Request data (Falcon 9 launch data) from Wikipedia
 - Create BeautifulSoup object from HTML response
 - Extract column names from HTML table header
 - Collect data from parsing HTML tables
 - Create dictionary from the data
 - Create dataframe from the dictionary
 - Export data to csv file
- 



Methodology

Data wrangling

The steps that were deemed necessary for data wrangling were:

- Perform EDA and determine data labels
- Calculate:
 - Quantity of launches for each site
 - Quantity and occurrence of orbit
 - Quantity and occurrence of mission
 - outcome per orbit type
- Create binary landing outcome column (dependent variable)
- Export data to csv file




Methodology

EDA

In this section, analyzes and charts were developed, namely:

- **Charts:**
 - Flight Number vs. Payload
 - Flight Number vs. Launch Site
 - Payload Mass (kg) vs. Launch Site
 - Payload Mass (kg) vs. Orbit type
- **Analysis:**
 - View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists
 - Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value.



Methodology

EDA with SQL

In this section, queries to understand about the data were developed:

- **Displayed:**
 - Names 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.
- **Listed:**
 - 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



Methodology

EDA with map Folium

To gain a more comprehensive understanding of the launch locations, maps were developed for visualization purposes:

- **Markers Indicating Launch Sites:**
 - Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
 - Added red circles at all launch sites coordinates with a popup label showing its name using its latitude and longitude coordinates
- **Colored Markers of Launch Outcomes:**
 - Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which launch sites have high success rates
- **Distances Between a Launch Site to Proximities:**
 - Added colored lines to show distance between launch site CCAFS SLC- 40 and its proximity to the nearest coastline, railway, highway, and city



Methodology

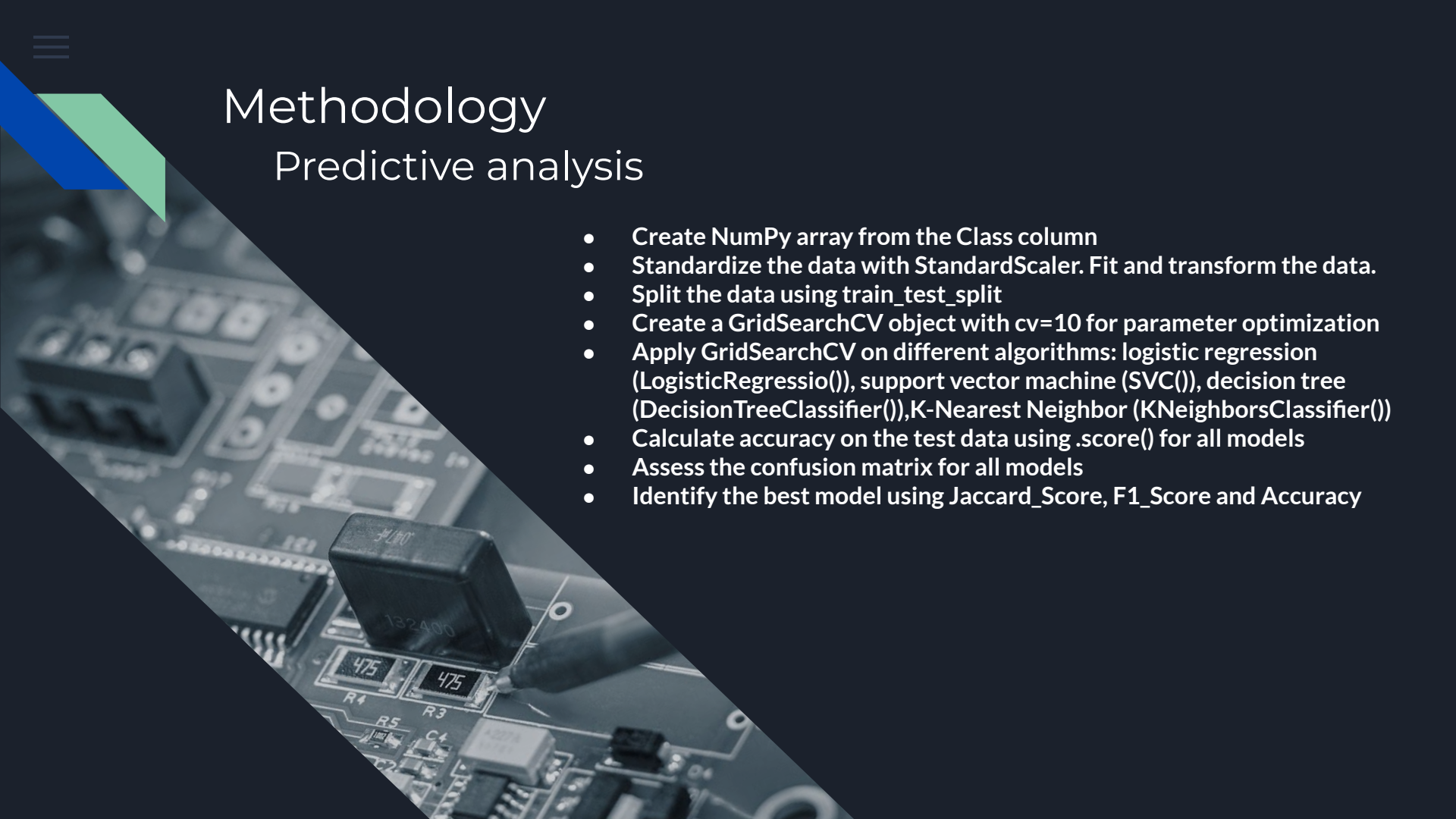
EDA with Plotly Dash

- **Dropdown List with Launch Sites**
 - Allow user to select all launch sites or a certain launch site
- **Pie Chart Showing Successful Launches**
 - Allow user to see successful and unsuccessful launches as a percent of the total
- **Slider of Payload Mass Range**
 - Allow user to select payload mass range
- **Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version**
 - Allow user to see the correlation between Payload and Launch Success



Methodology

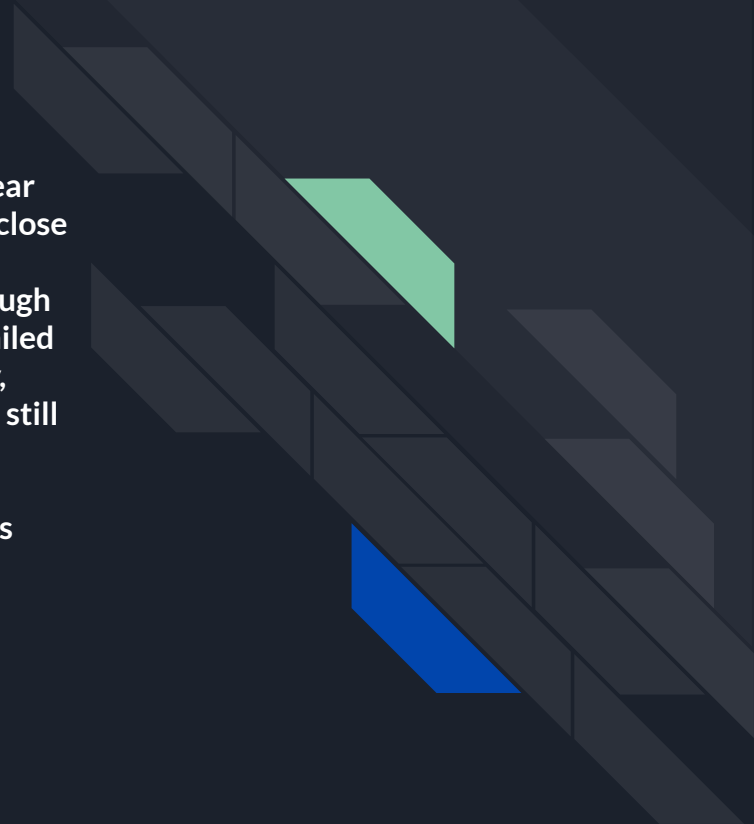
Predictive analysis

- 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

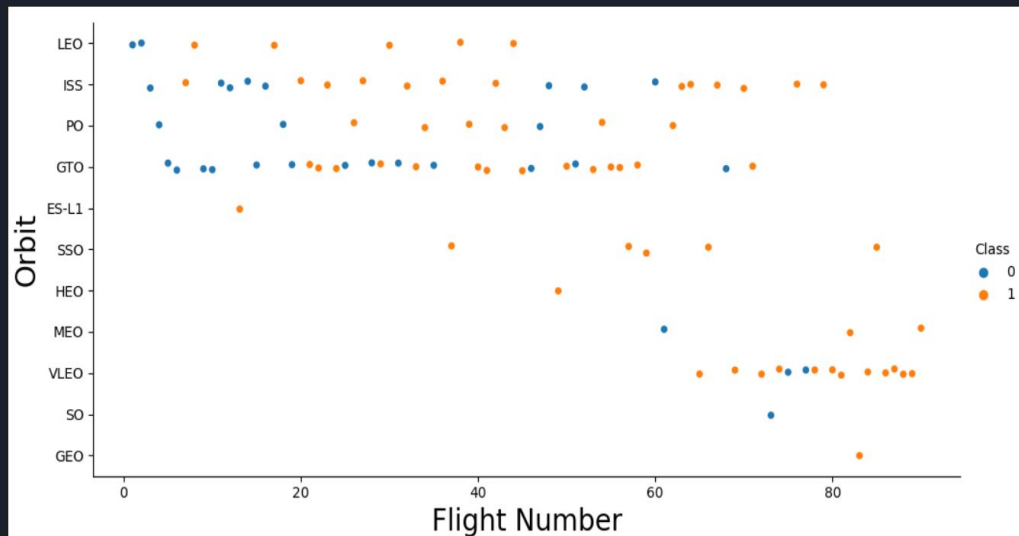
Results summary:

- **Exploratory Data Analysis:**
 - Launch success has improved over time
 - KSC LC-39A has the highest success rate among landing sites
 - Orbits ES-L1, GEO, HEO and SSO have a 100% success rate
- **Predictive Analytics:**
 - Decision Tree model is the best predictive model for the dataset
- **Visual Analytics:**
 - Most launch sites are near the equator, and all are close to the coast
 - Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities



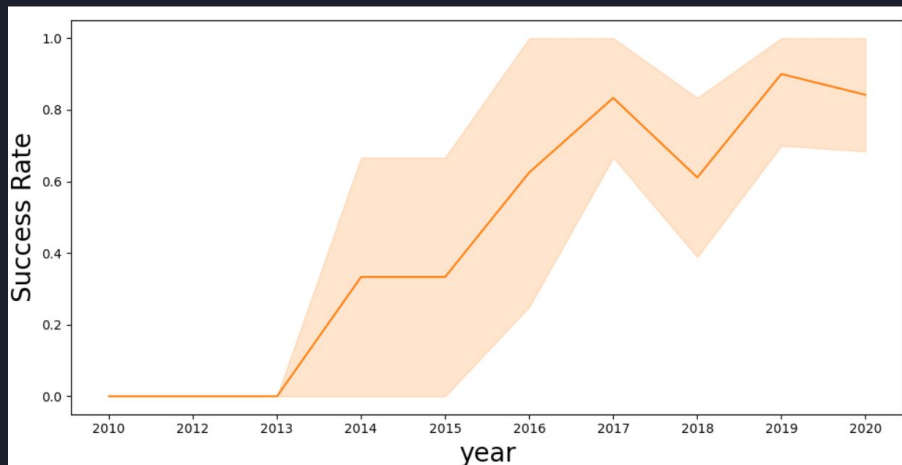
Results

- **Flight Number vs. Launch Site:**
 - VAFB SLC 4E and KSC LC 39A have higher success rates
 - We can infer that new launches have a higher success rate



Results

- Launch Success over Time:
 - The success rate improved from 2013-2017 and 2018-2019
 - The success rate decreased from 2017-2018 and from 2019-2020
 - Overall, the success rate has improved since 2013



Results

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

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) \
      FROM SPACEXTBL \
      WHERE CUSTOMER = 'NASA (CRS)';
```

SUM(PAYLOAD_MASS__KG_)

45596

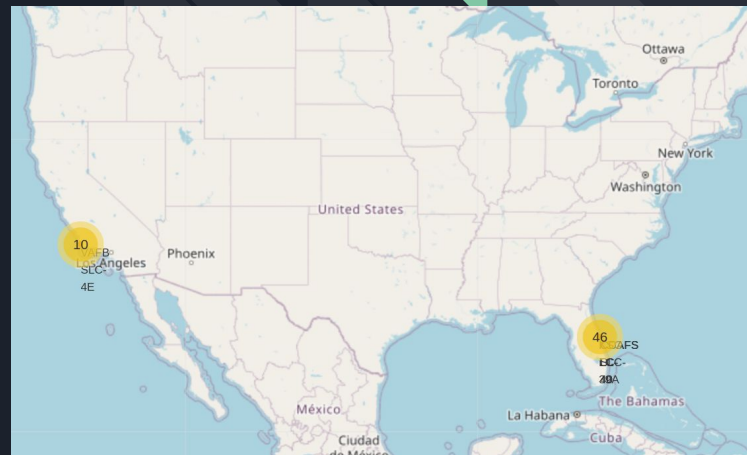
Results

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

```
%sql SELECT AVG(PAYLOAD_MASS_KG_) \
FROM SPACEXTBL \
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

AVG(PAYLOAD_MASS_KG_)

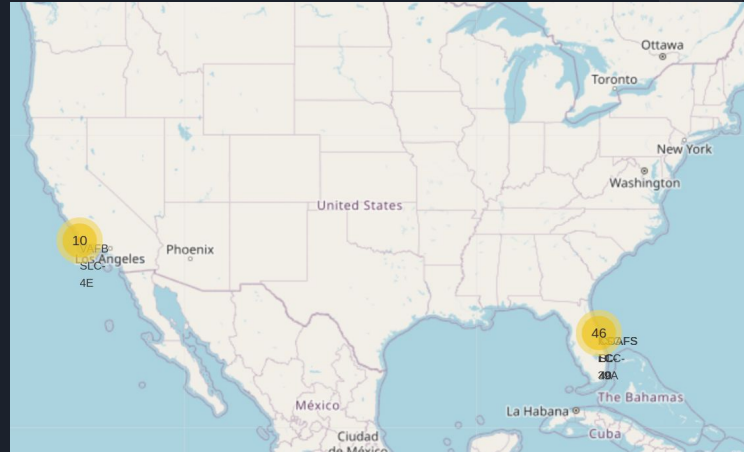
2928.4



Results

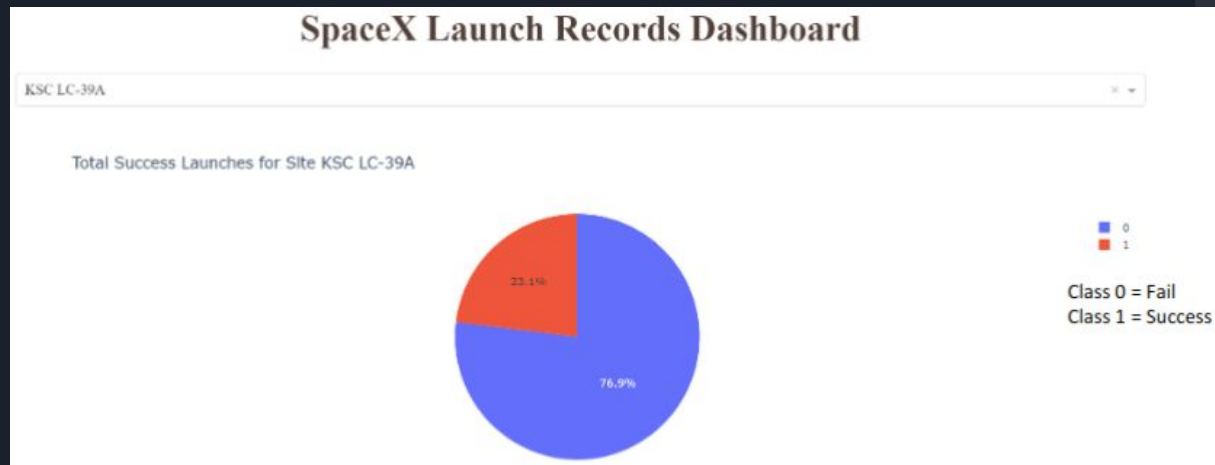
- Launch Sites

- Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



Results

- Launch Success (KSC LC-29A)
 - KSC LC-39A has the highest success rate amongst launch sites (76.9%)
 - 10 successful launches and 3 failed launches



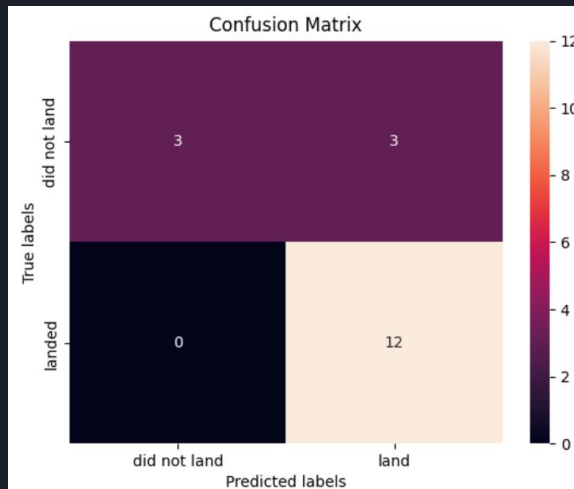
Results

- Predictive analysis
 - All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset. The Decision Tree model slightly outperformed the rest when looking at .best_score_

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

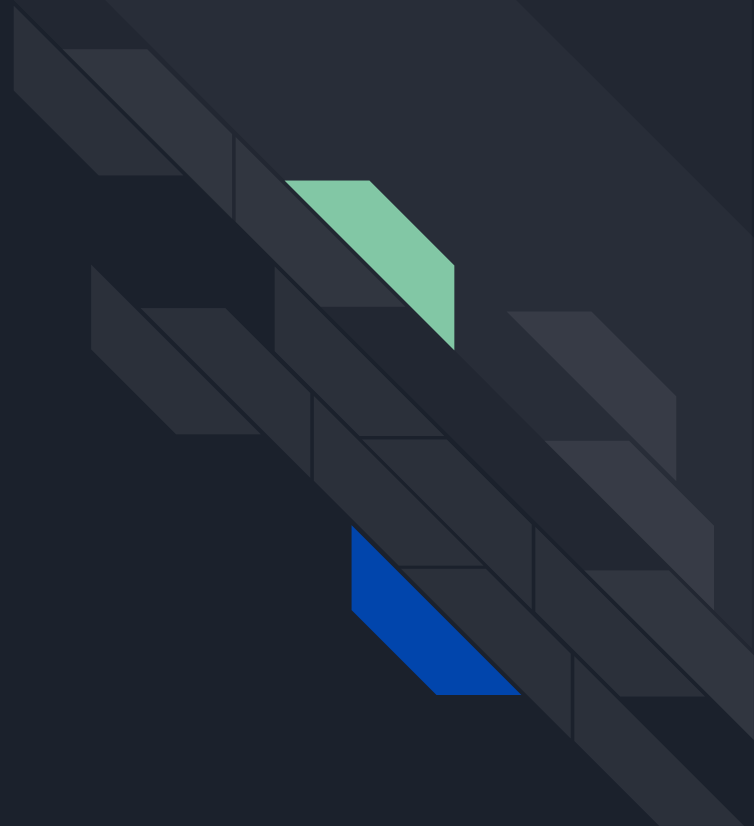
Results

- Predictive analysis
 - Performance Summary:
 - A confusion matrix summarizes the performance of a classification algorithm
 - Confusion Matrix Outputs:
 - 12 True positive
 - 3 True negative
 - 3 False positive
 - 0 False Negative



Conclusion

- **Model Performance:** The models performed similarly on the test set with the decision tree model slightly outperforming
- **Coast:** All the launch sites are close to the coast
- **Launch Success:** Increases over time
- **KSC LC-39A:** Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- **Dataset:** A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set





Thank you!

Until next time!