IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Akhil Srirangam
7/5/2024

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Data Visualization
  - Building Interactive Maps and Interfaces with Folium and Dash
  - Predictive Analysis using Scikit-Learn
- Summary of all results
  - Exploratory Data Analysis results
  - Folium and Dash Screenshots for Interactive Maps and Interfaces
  - Results with accuracy for Predictive Analysis

Introduction

- Project background and context
  - In this project, we took on the role of a data scientist aiming to help SpaceX. When rockets are launched, the first stage of a rocket is the most expensive, and SpaceX found a way to reduce it. Our mission was to predict if, given different factors, if SpaceX could successfully land and reuse the first stage of the rocket.
- Problems you want to find answers
  - How do different features, such as launch site, number of flights, and orbit affect the success of landing the first stage?
  - Does the year launch affect success rate?
  - In this case, through binary (categorical) classification, what is the best model we can use to predict the success?

Section 1

Methodology

Methodology

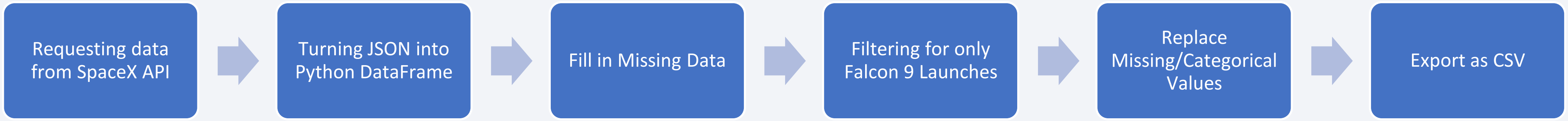
Executive Summary

- Data collection methodology:
  - Using the SpaceX Rest API
  - Using Web Scraping from Wikipedia
- Perform data wrangling
  - Filtering missing values
  - Using one-hot encoding to prepare the data for binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building models, and evaluating and tuning for best accuracy

Data Collection

- Data Collection
  - Using the SpaceX Rest API – included all of the date in a ready-to-use format, but had some missing values found in other APIs
  - Using Web Scraping from Wikipedia – used this to have more data not found/better formatted than in the API

Data Collection – SpaceX API



[GitHub URL to Data Collection using API]

Data Collection - Scraping

Requesting launch data from Wikipedia → Using BeautifulSoup to begin webscraping → Extracting data from tables using HTML notation → Constructing data into a dictionary → Creating a DataFrame from dictionary → Export as CSV

[GitHub URL to Data Collection using Wikipedia]()

Data Wrangling

- Some of the categories, such as landing success/failure was a categorical value, with 5+ different outcomes. Other categories that had categorical features include orbits and rocket type.
- The way we dealt with this is one-hot encoding, which makes a new column for each possible outcome, and puts a 1 to indicate which kind it is, and a 0 everywhere else.
- This helps binary classifier models develop a more accurate algorithms to predict the correct result

GitHub URL to Data Wrangling

EDA with Data Visualization

- We used three main types of charts:
  - Scatter plots – shows the relationship between two variables; these relationships can be used by machine learning models
  - Bar Charts show comparisons between categories; these relationships can help models make decisions with more certainty
  - Line charts show the progression of data over time; these relationships can uncover trends based on time

GitHub URL to EDA with Visualization

EDA with SQL

- Using SQL, we performed various queries that displayed unique launch sites, payload mass details, gaining data about successful launch landings, dividing successful landings by payload mass, and gaining data about failed launch landings

[GitHub URL to EDA with SQL]

Build an Interactive Map with Folium

- Folium is a library that helps us make an interactive map
- We used it to make marker objects of the launch sites, which gives us the launch sites longitude and latitude coordinates, as well as proximity to the Equator and relevant coasts
- The marker objects were color coded based on their success (Green/Red)
- The marker objects also gave us proximity to relevant utilities such as railways, highways, etc.

GitHub URL to Folium

Build a Dashboard with Plotly Dash

- Consisted of:
  - Dropdown list for launch site selection
  - Pie chart to show successful launches, whether it is all sites or a selected site
  - Had a slider of payload mass range to filter results
  - Scatter chart to show relation between mission success and payload mass

GitHub URL to Dashboard

Predictive Analysis (Classification)

Creating NumPy array from data and scaling it using StandardScaler

→

Split data into training and testing sets using train_test_split

→

Using GridSearchCV to optimize hyperparameters

→

Use GridSearchCV on LogReg, SVM, DecisionTrees, and KNN ML Classifier Models

→

Calculating accuracy using .score() and metrics, constructing confusion matrix, and choosing highest accuracy

GitHub URL to Predictive Analysis

Results

EXPLORATORY DATA ANALYSIS RESULTS

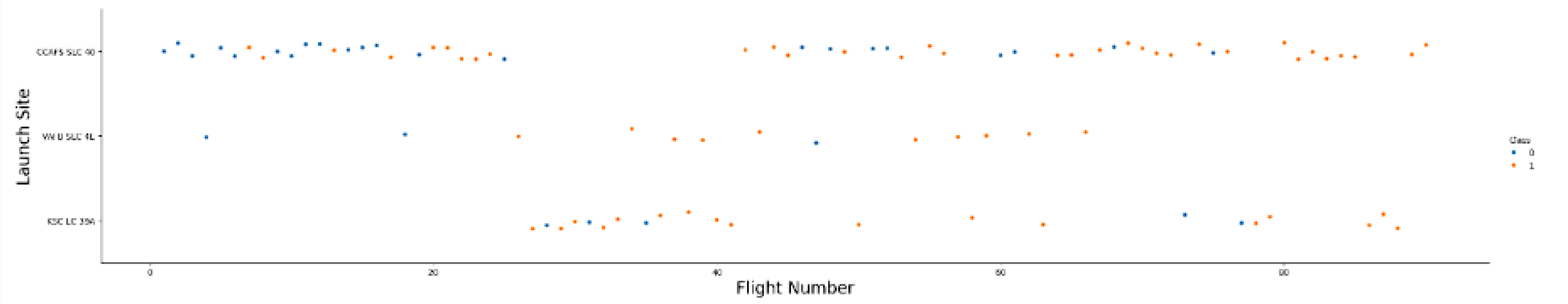
INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

PREDICTIVE ANALYSIS RESULTS

Section 2

Insights drawn from EDA

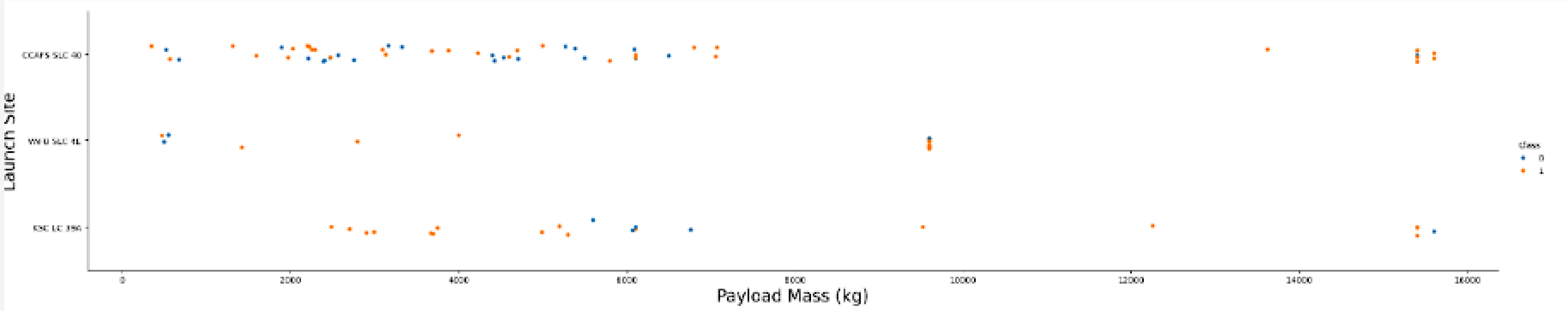
Flight Number vs. Launch Site



Analysis:

- CCAFS SLC has about half of all launches
- The other two launch sites have a higher success rate
- Newer launches show more likelihood of success

Payload vs. Launch Site

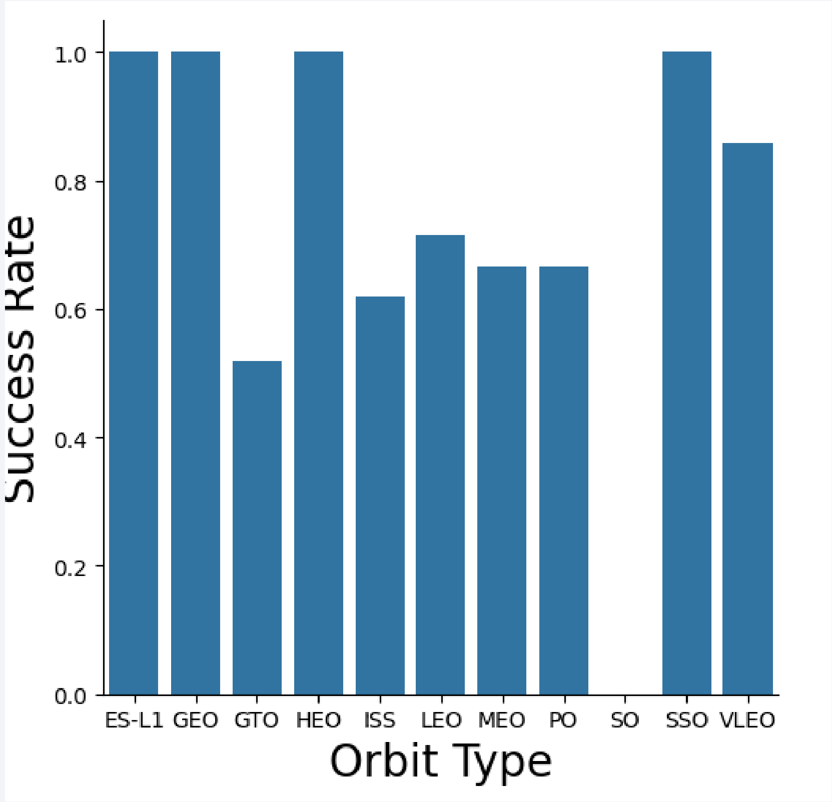


Analysis:

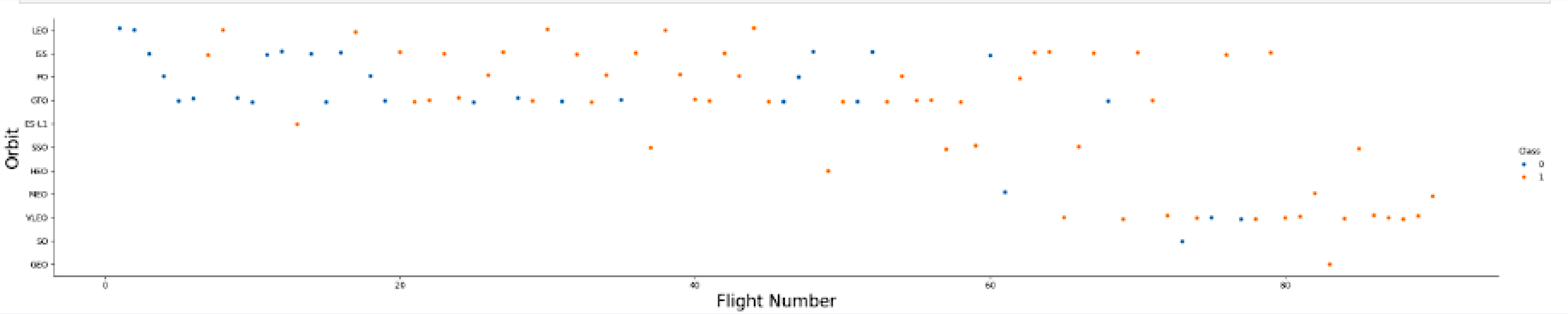
- Higher the payload mass, higher the success rate
- Most launches with payload mass over 7000 kg were successful
- KSC LC 39A has a 100% success rate with mass under 5500kg

Success Rate vs. Orbit Type

- 100% success rate groups are visible – helpful for models
- 0% success rate group is visible – helpful for models



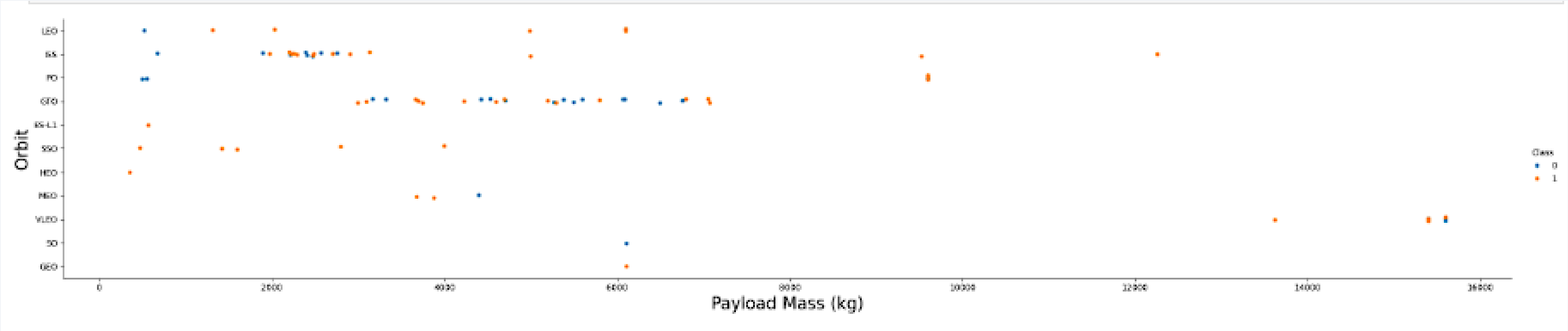
Flight Number vs. Orbit Type



Analysis:

- No apparent correlation between flight number and orbit

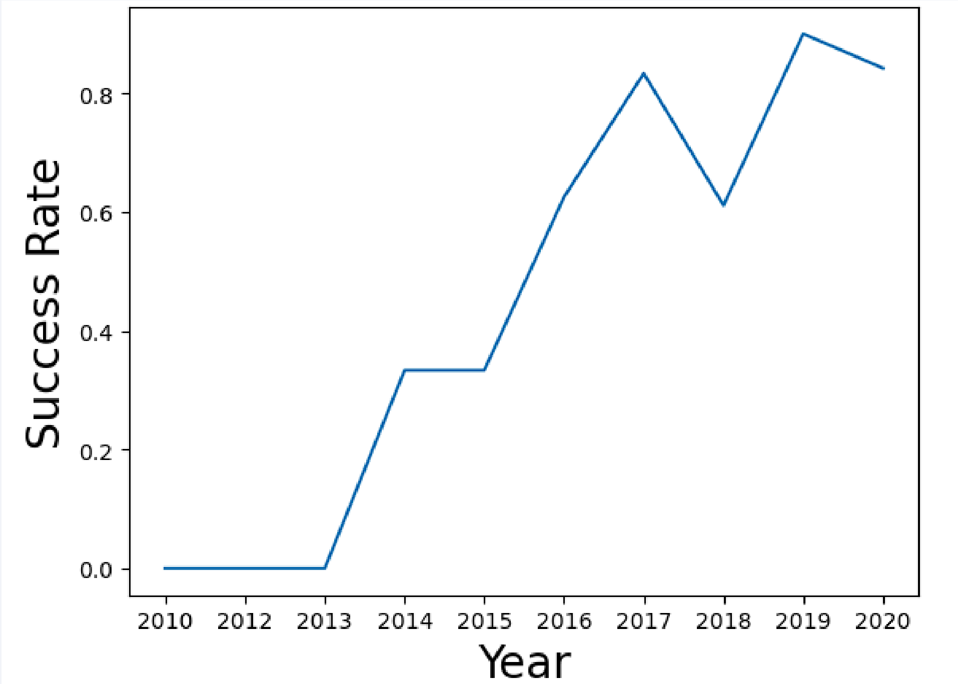
Payload vs. Orbit Type



Analysis:

- Heavy payloads affect different orbit types differently
  - GTO has negative influence and ISS has positive influence etc.

Launch Success Yearly Trend



Analysis:

- Success rate increases 2013-2017, and again 2018-2019

All Launch Site Names

- This query displays the unique launch sites from the dataset

```
%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'

- This query returns 5 records of launches from the CCAFS LC-40 launch site

```
%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_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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 (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- This query returns the total payload mass derived from NASA

```
%sql SELECT sum(payload_mass__kg_) AS total_payload_mass FROM SPACEXTABLE WHERE customer = 'NASA (CRS)';
```

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- This query returns the average payload mass of rockets that used the Falcon 9 version 1.1 booster

```
%sql SELECT avg(payload_mass__kg_) AS average_payload_mass FROM SPACEXTABLE WHERE booster_version LIKE '%F9 v1.1%';
```

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

average_payload_mass
2534.6666666666665

First Successful Ground Landing Date

- This query shows the date of the first successful ground pad landing

```
%sql SELECT min(date) AS first_successful_landing FROM SPACEXTABLE WHERE landing_outcome = 'Success (ground pad)';
```

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

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- This query shows the booster versions of rockets that had a successful drone ship landing, with payload mass between 4000 and 6000 kg

```
%sql SELECT booster_version FROM SPACEXTABLE WHERE landing_outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 400(
```

```
* 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

- This query returns the number of successful and failure missions. The reason there are multiple success columns is because there are different types of successes in the records

```
%sql SELECT mission_outcome, COUNT(*) AS total_number FROM SPACEXTABLE GROUP BY mission_outcome;
```

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

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

Boosters Carried Maximum Payload

- This query returns the booster versions of the Falcon 9 rockets that were involved in carrying the maximum recorded payload mass

```
%sql SELECT booster_version FROM SPACEXTABLE WHERE payload_mass__kg_ = (SELECT max(payload_mass__kg_) FROM SPACEXTABLE);
```

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

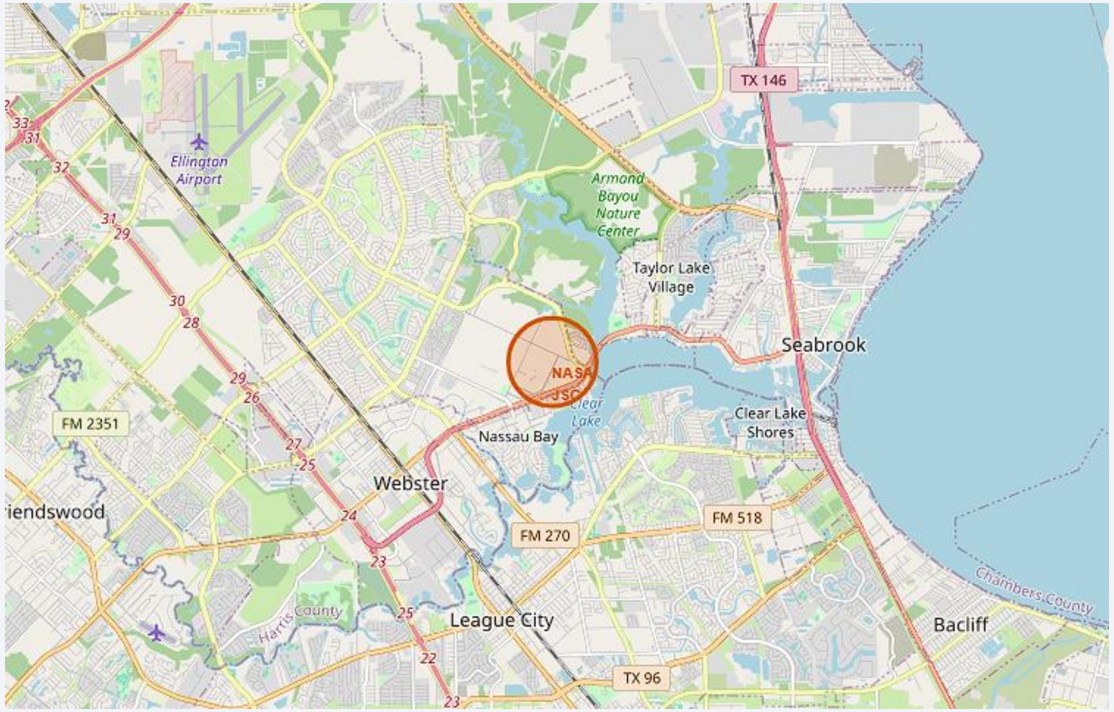
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Section 3

Launch Sites Proximities Analysis

Folium NASA Marker

- We used Markers such as this to mark different launch sites and NASA

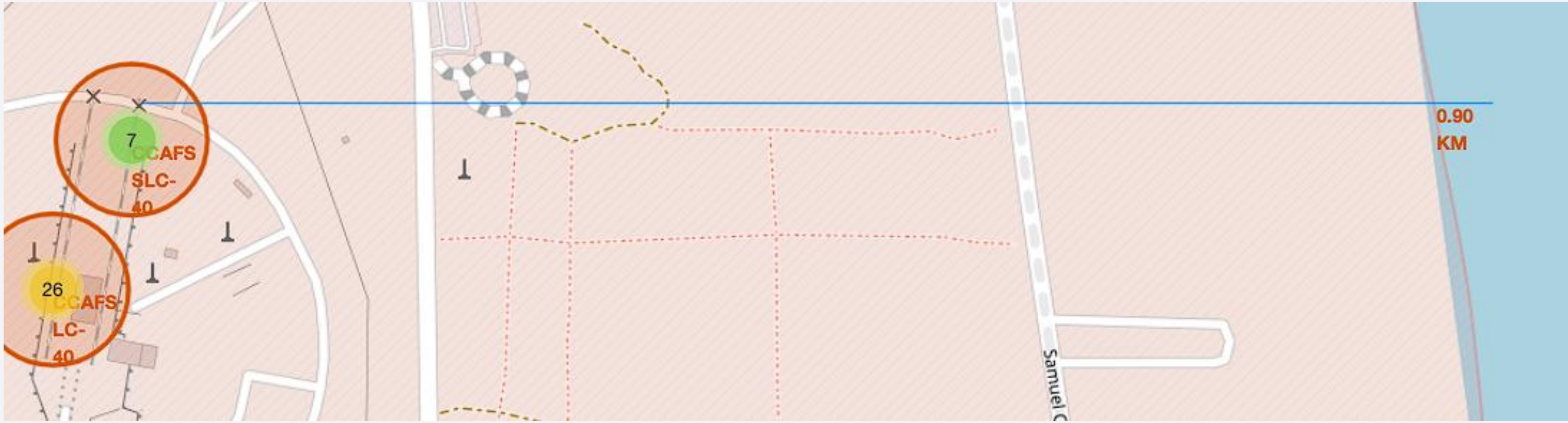


Color Coding Folium Markers

- The markers were color coded to show successful launches. The screenshot shows the results for one site, but this schematic was replicated across all launch sites. The color-coded markers have icons that show launch information.

Folium Map with Proximity to Features

- The map now has a feature where it can show the proximity to features such as highways or coastlines

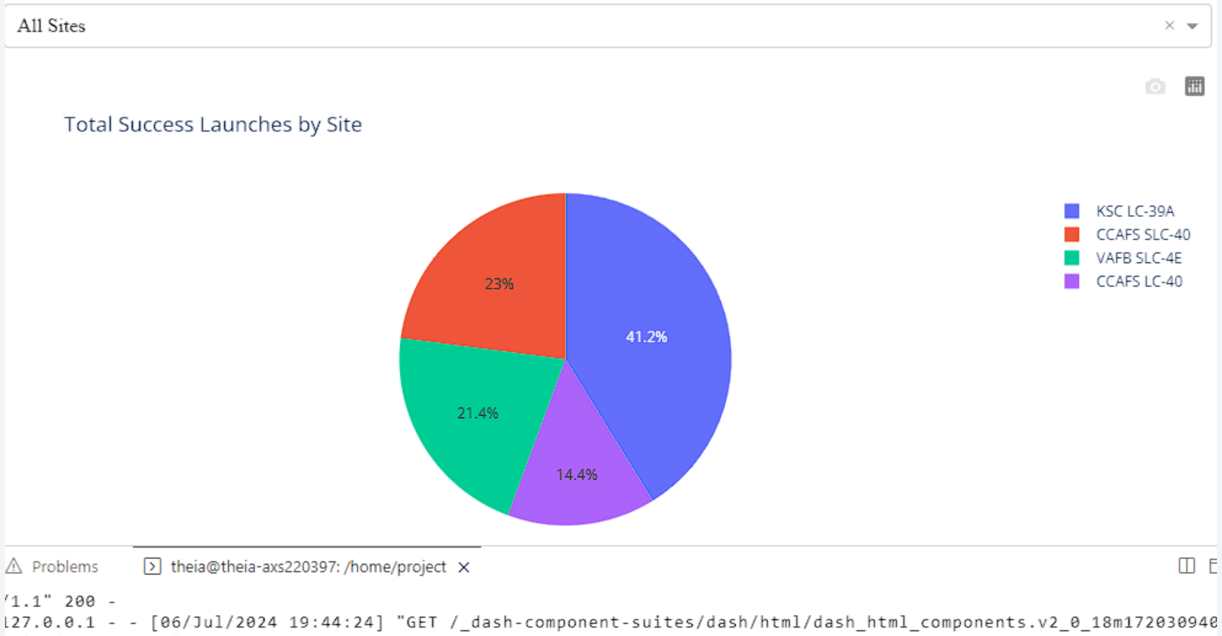


Section 4

Build a Dashboard with Plotly Dash

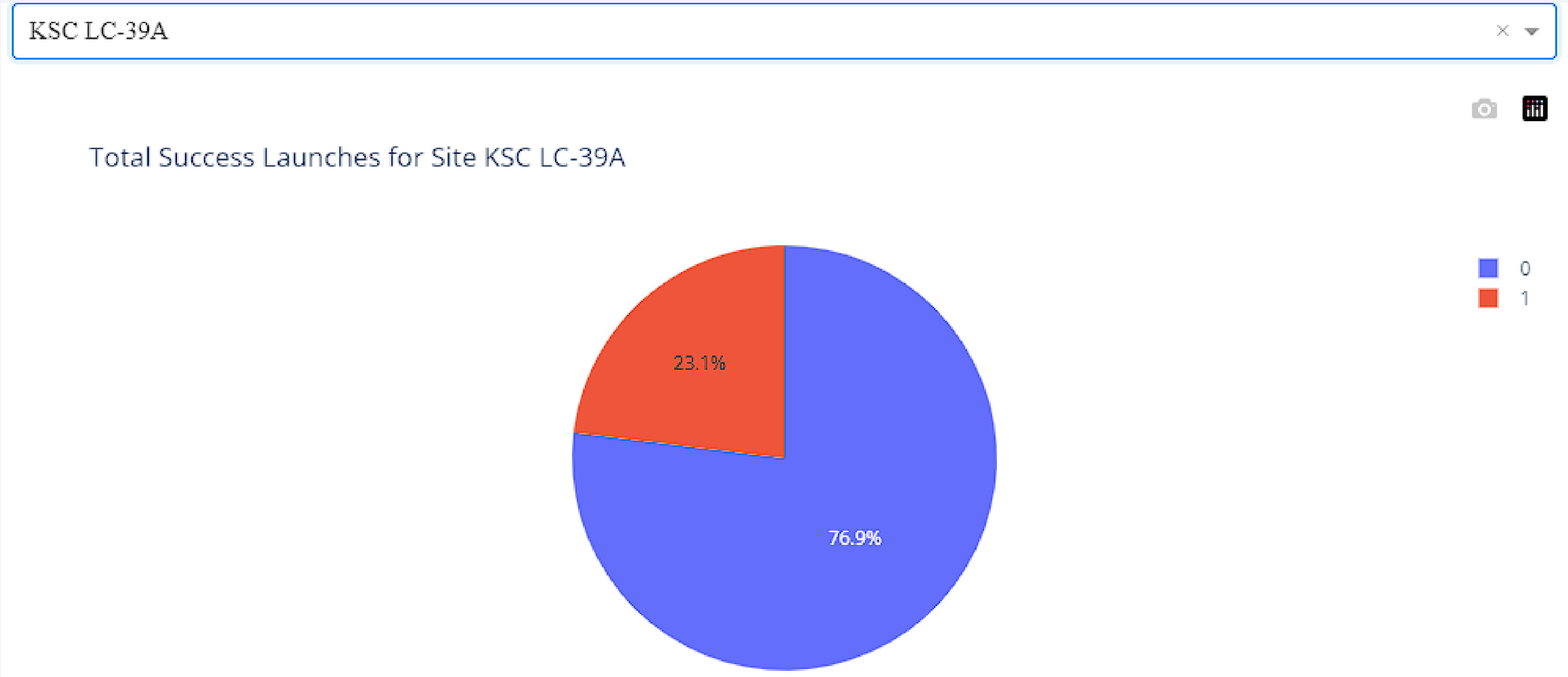
Launch Success Count for All Sites

- This graphic displays a pie chart that shows how many successful landings belong to each landing site



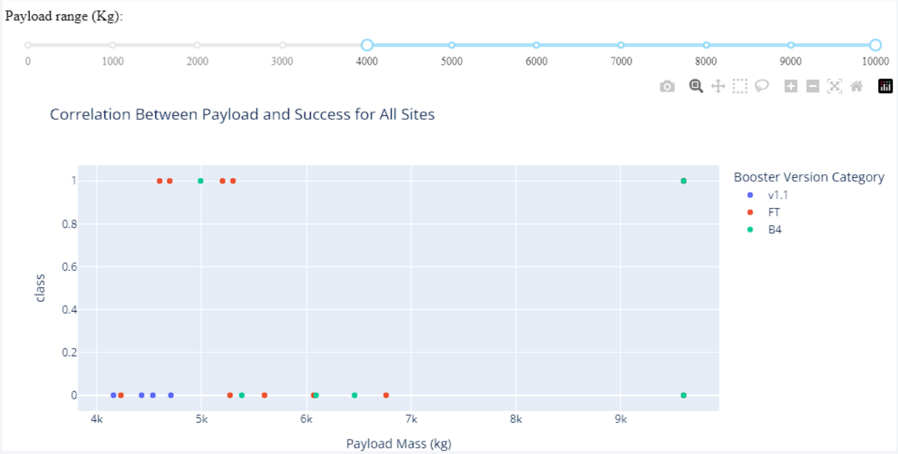
Success Graphic for Site with Highest Success

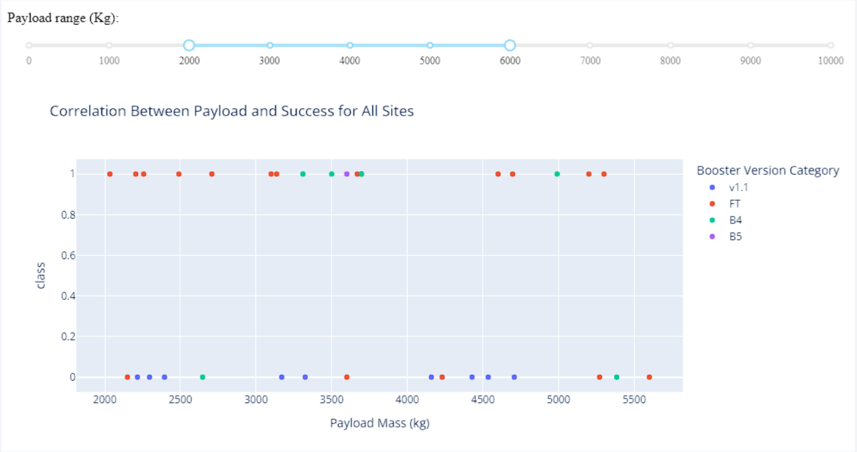
- This graphic shows the success rate of the site with the highest successful landing ratio, KSC LC 39-A



Payload vs Launch Outcome Plots w/ Different Payload Weights

- These graphics show the Scatter Plots displaying Payload Weight vs Outcome relationships
- These show the relationship between booster versions and success ratios by weight



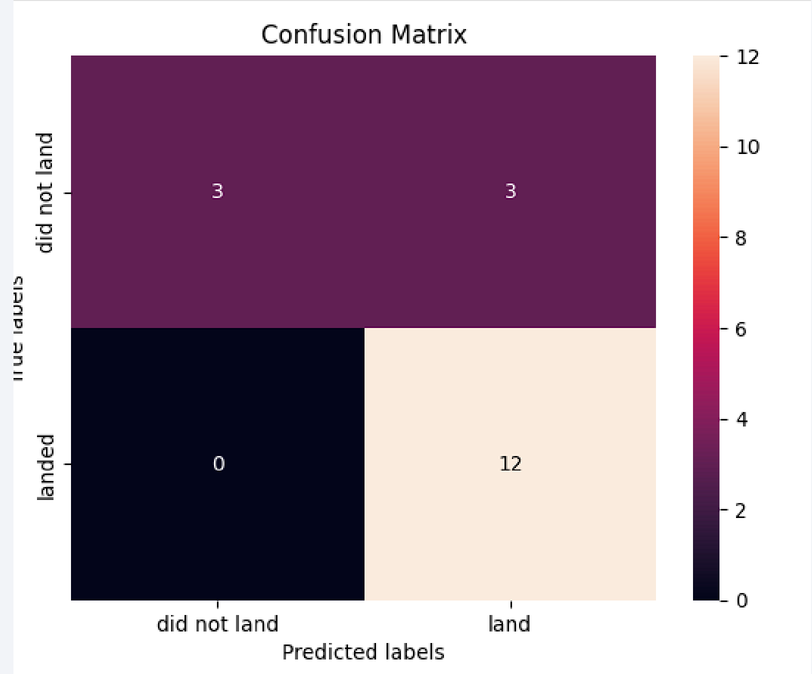


Section 5

Predictive Analysis (Classification)

Confusion Matrix

- Below is the Confusion Matrix of the best performing Machine Learning Model, the SVM (Support Vector Machine
- This means that there were:
  - 3 False Successes
  - 3 True Failures
  - 0 False Failures
  - 12 True Successes



Conclusions

- Some categories such as launch site, payload weight, and which orbit it is entering have a massive impact on the success of a landing of the 1^{st} stage of the rocket booster
- The rate of successful landings, on average, increased over time
- The best algorithm that can be used as a binary classifier is the SVM (Support Vector Machine), because of its higher scores in metrics

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.666667	0.819444
F1_Score	0.909091	0.916031	0.800000	0.900763
Accuracy	0.866667	0.877778	0.666667	0.855556

Thank you!