IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Andy Hadian Shah
25th October 2025

Outline

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

Executive Summary

Summary of methodologies

- Data Collection from APIs and Web Scraping
- Data Wrangling
- EDA with Data Visualization and SQL
- Interactive Maps and Dashboards
- Predictive Analysis with Machine Learning

Summary of all results

- Decision Tree Classifier outperformed the other tested classifiers
- The classifier was able to predict the outcome of a mission to be either successful or failure consistently with lower precision

Introduction

- SpaceX spends around $62 million to conduct their space launches while other provides spend around $165 million. SpaceX can do so by reusing their first stage of the rocket. By being able to predict whether a rockets first stage can be reused and determine the price of a mission predictively, which can be used for mission biddings against SpaceX or help SpaceX with improving the chances of their launches being successful.
- To be able to do this, we need to know where can we collect data of SpaceX launches, what are the hidden patterns between columns of the dataset and what classifier is the best to predict whether a mission is successful or a failure.

Section 1

Methodology

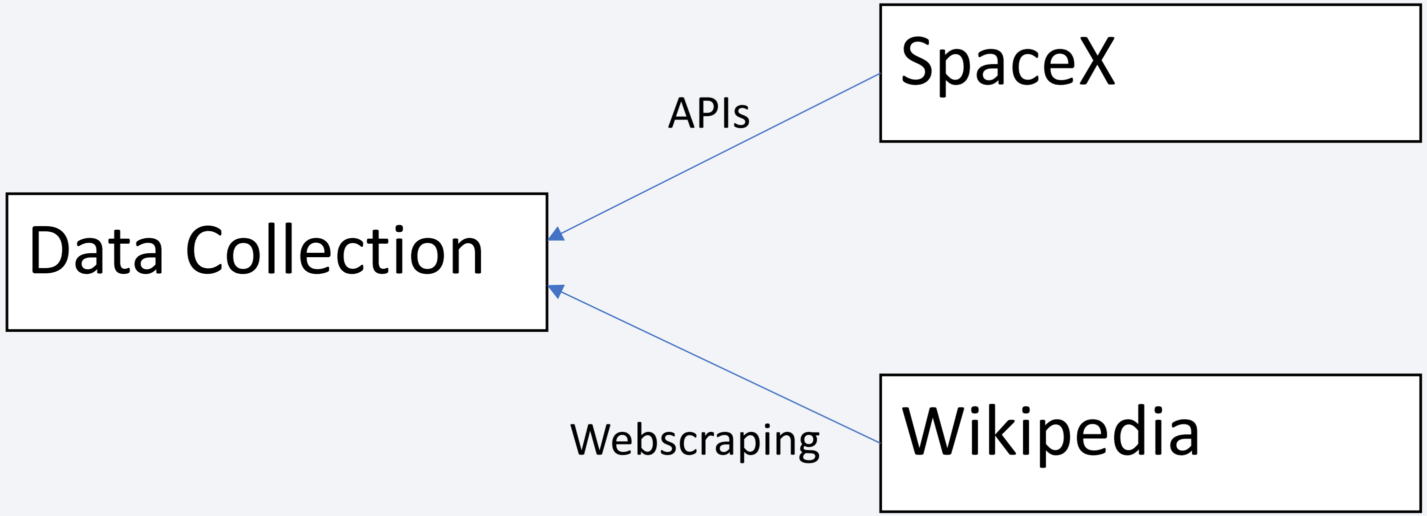
Methodology

Executive Summary

- Data collection methodology:
  - Using APIs and Webscraping from appropriate sources
- Perform data wrangling
  - Filtering for only Falcon 9 Launches, Removing missing data, analysis of categorical columns, and creating class label column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Dataset is prepared for training and testing models, multiple classifiers are trained and evaluated accordingly and the best forming model in accurately predicting mission outcomes is picked to be used for predictive analysis.

Data Collection

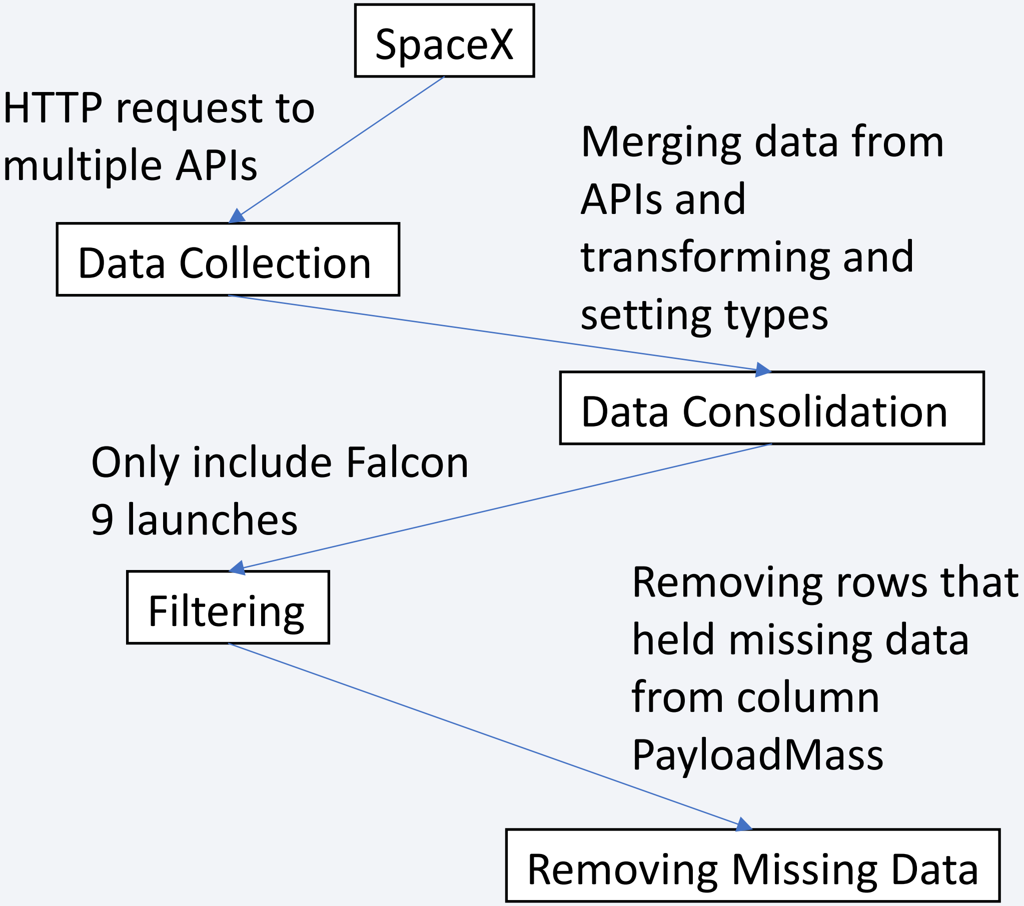
- The datasets used in this analysis made use of APIs to extract data directly from related sources, specifically from SpaceX's APIs, while also using web scraping to extract launch records from sources like Wikipedia tables.



Data Collection – SpaceX API

- Data was given from the SpaceX APIs. They were then merged into a single dataset where only Falcon 9 launches were kept. The data was then wrangled to remove missing data from PayloadMass
- Flowchart of the process can be seen in the diagram to the right

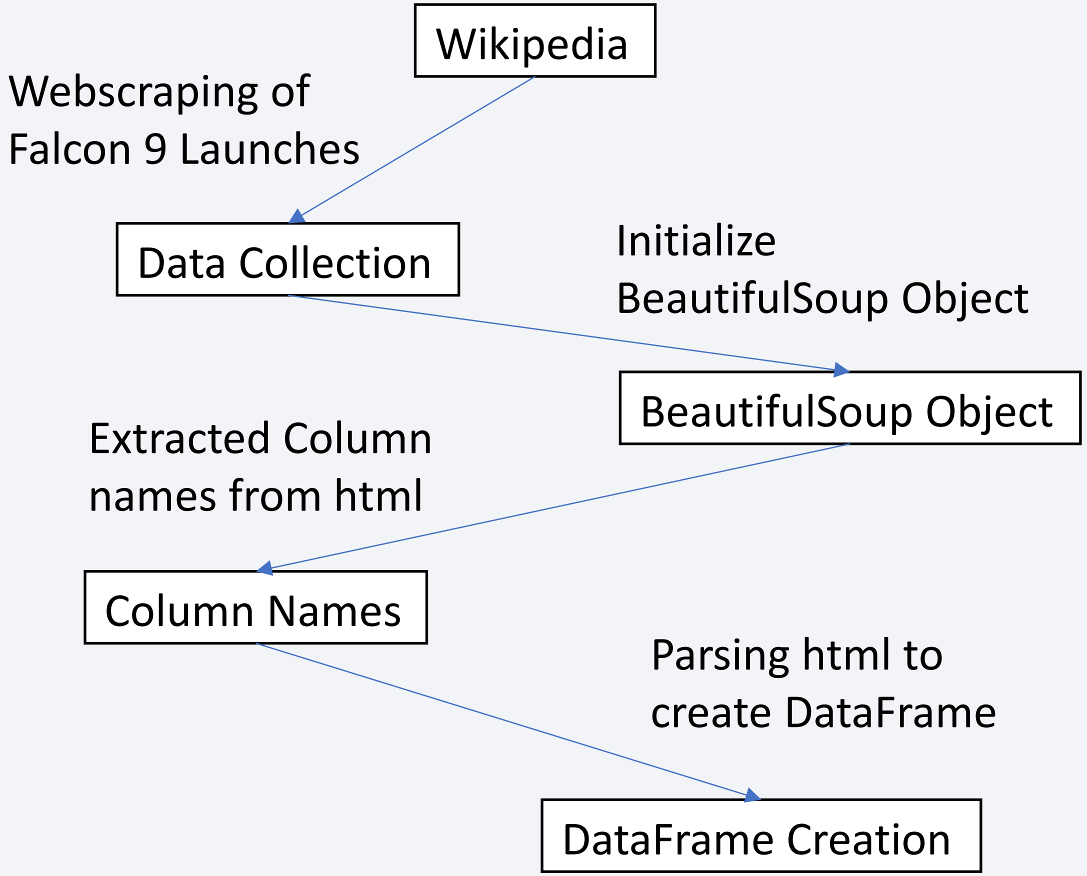
GitHub URL:
https://github.com/Swasbuckler/DataScienceAppliedCapstone/blob/main/jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Data was scrapped from the Wikipedia of Falcon 9 Launches. BeautifulSoup was used to parse through the html and extract data from the launches table and saved in a dataset.
- Flowchart of the process can be seen in the diagram to the right

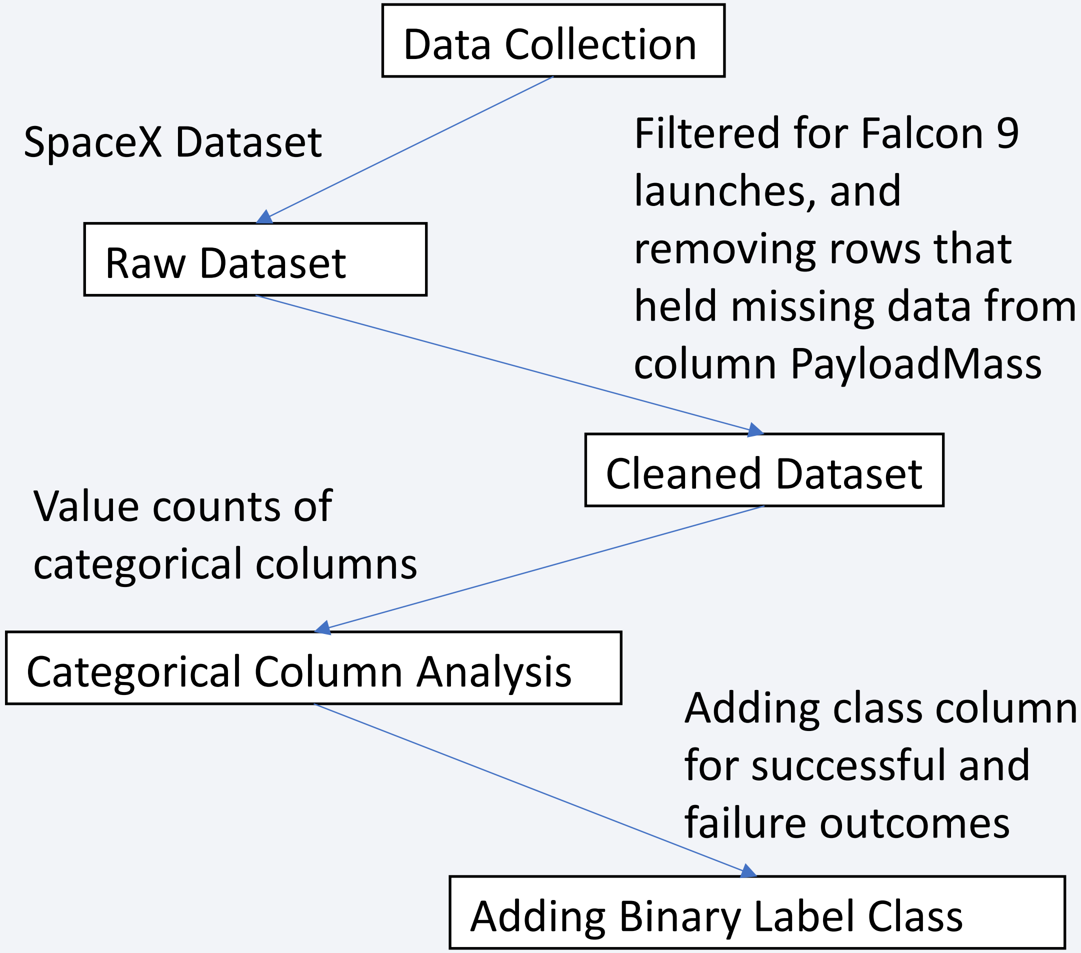
GitHub URL:
https://github.com/Swasbuckler/DataScienceAppliedCapstone/blob/main/jupyter-labs-webscraping.ipynb



Data Wrangling

- Data from the SpaceX APIs were taken, which had previously been filtered for only containing Falcon 9 launches and cleaned of missing values. The data was then analyzed of the value counts of categorical columns like Launch Sites, Orbits and Outcomes. A label binary class was then added to the dataset named “class” for successful and failure outcomes.
- Flowchart of the process can be seen in the diagram to the right

GitHub URL:
https://github.com/Swasbuckler/DataScienceAppliedCapstone/blob/main/labs-jupyter-spacex-Data%20wrangling.ipynb



EDA with Data Visualization

- The Charts used for Data Visualization include Catplots, Scatter Plots, Bar Plot and Line Plot.
- Catplots, Scatter Plots and Bar Plot were chosen to help show the relationship between categorical and numeric columns. Doing so helped with identifying the hidden patterns found between columns like Orbits, Launch sites, Flight numbers, Payload mass etc.
- Line Plot was used to learn of time series patterns through the years.

GitHub URL:
https://github.com/Swasbuckler/DataScienceAppliedCapstone/blob/main/edadataviz.ipynb

EDA with SQL

SQL queries that have been used to extract data from the dataset which includes,

- Unique data of columns
- Extract specific rows based on conditions such as Launch site starting with a substring, first ever launch date, etc.
- Using aggregate functions to determine average, count, sum, min and max values

GitHub URL:
https://github.com/Swasbuckler/DataScienceAppliedCapstone/blob/main/jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

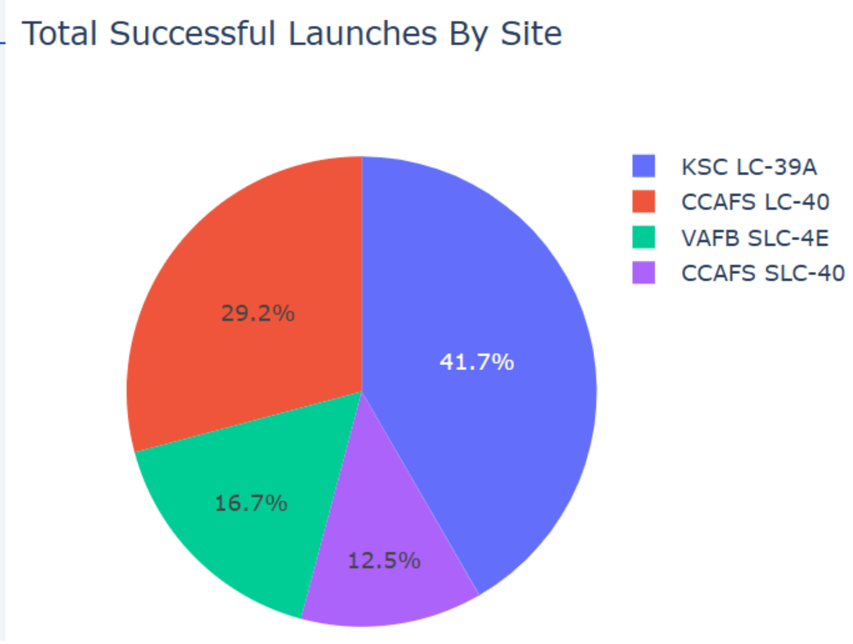
- Iterative maps were used to position launch sites across the earth to compare where they are from one another, whilst also seeing their proximity to other locations on earth, like with other cities and roads.
- Markers, Labels, Circles and Lines help with positioning launch sites on the map no matter how far the maps are zoomed out for better a user experience. Lines and labels helped to identify what is being seen on the map and view distances between locations.

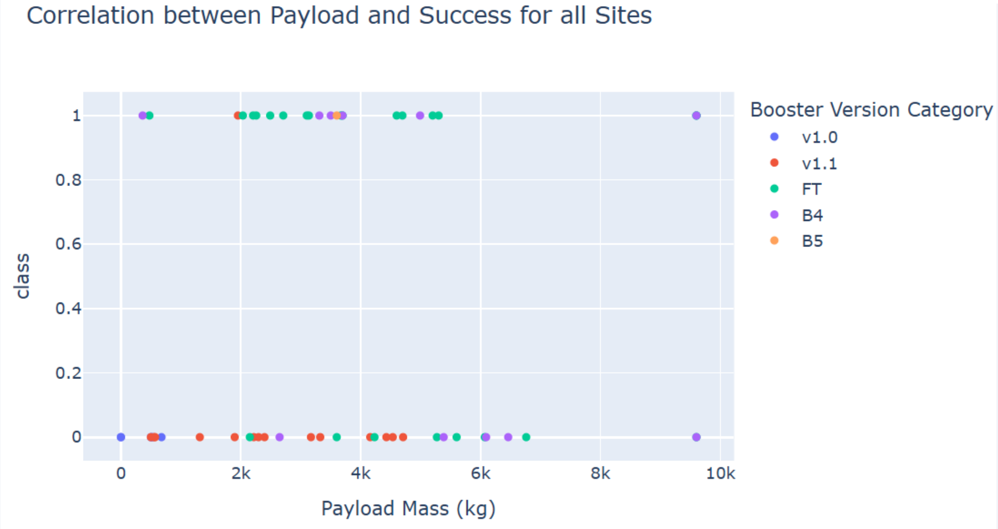
GitHub URL:

https://github.com/Swasbuckler/DataScienceAppliedCapstone/blob/main/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie charts and scatter plots were shown where users can interact with them through a dropdown and range selector.
- For the dropdown, users can select to view data on all launch sites or a specific site. While the range selector selects on the scatter plot mass.
- These interactivity helps user to make a



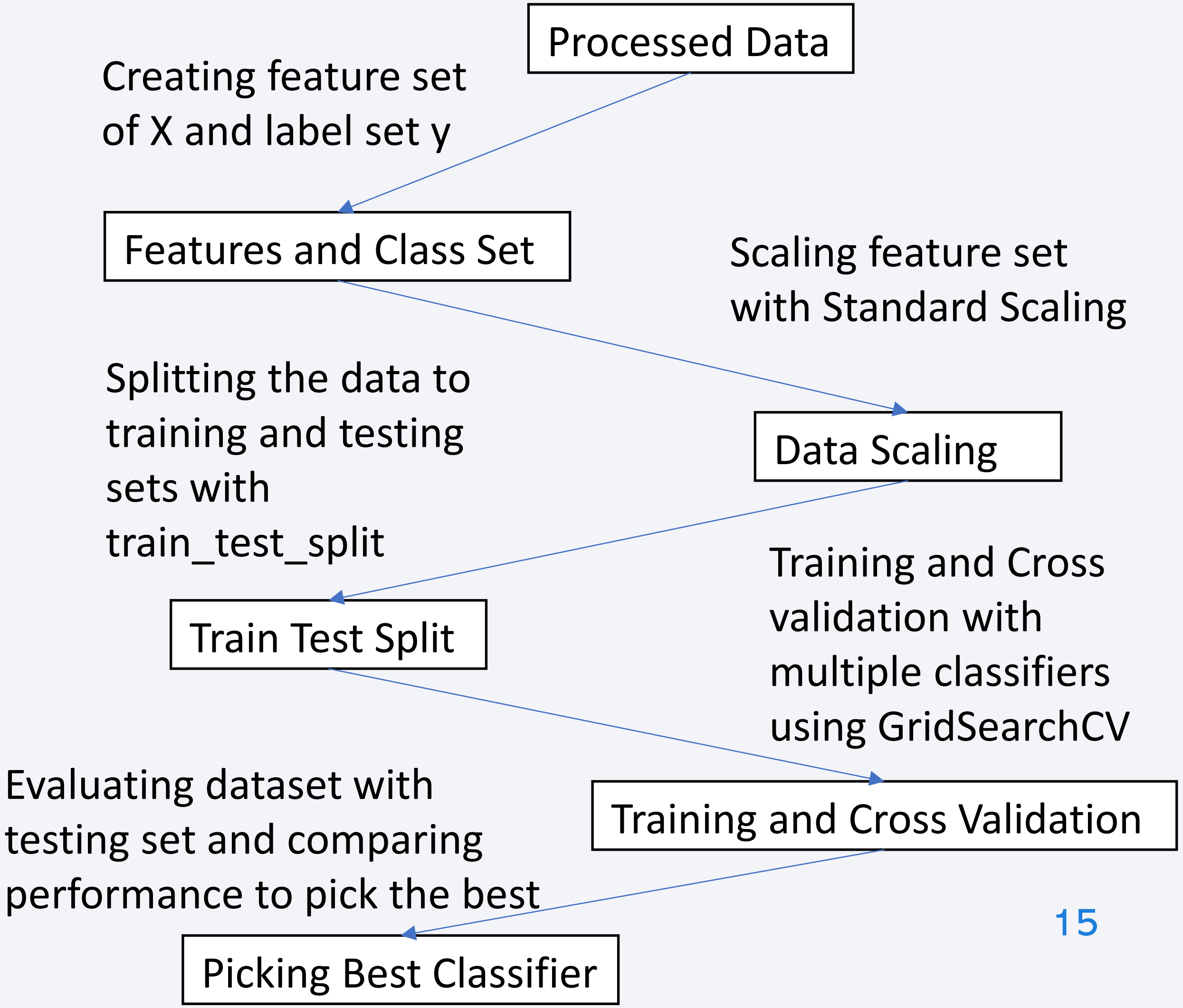


GitHub URL: https://github.com/Swasbuckler/DataScienceAppliedCapstone/blob/main/spacex-dash-app.py

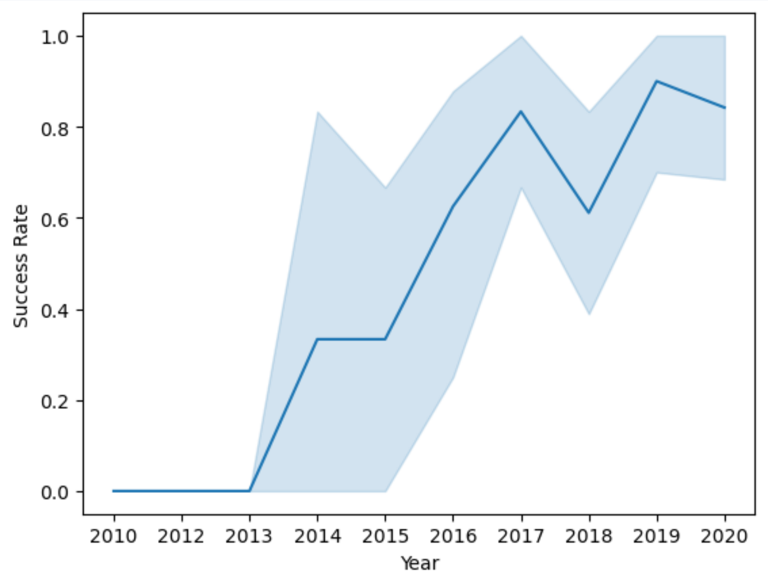
Predictive Analysis (Classification)

- The processed data after data wrangling is separated to a feature and label sets. The feature set is then scaled and the whole dataset is split into a train and test set with 20% going to the testing set. GridSearchCV is used to do Cross Validation of 10 and find the best parameters for classifiers Logistic Regression, Support Vector Classifier, Decision Tree and K-Nearest Neighbour. The classifiers are evaluated on accuracy and the best performing one is picked.
- Flowchart of the process can be seen in the diagram to the right

GitHub URL:
https://github.com/Swasbuckler/DataScienceAppliedCapstone/blob/main/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

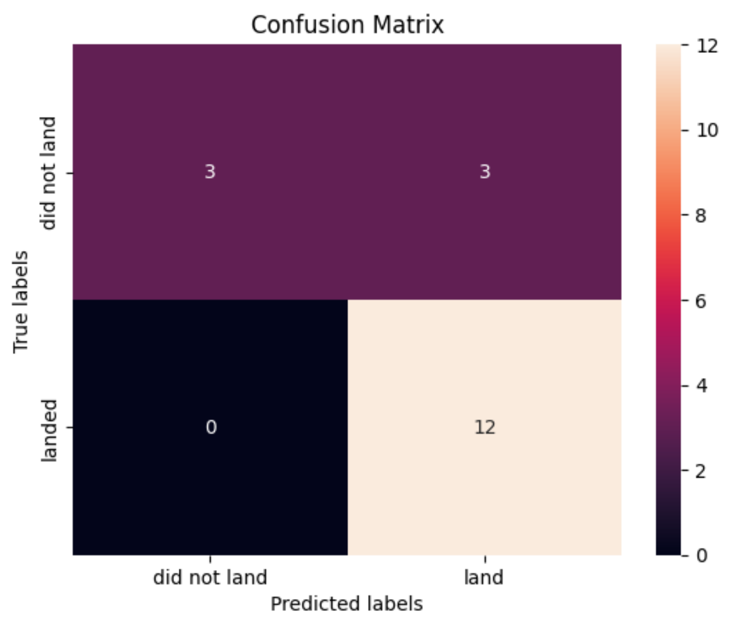


Results



Interactive analytics:

- Launch sites placed on an interactive map
- Sites are near coasts and away from human settlements



EDA analysis:

- Over the years mission success rates increases

Predictive analytics:

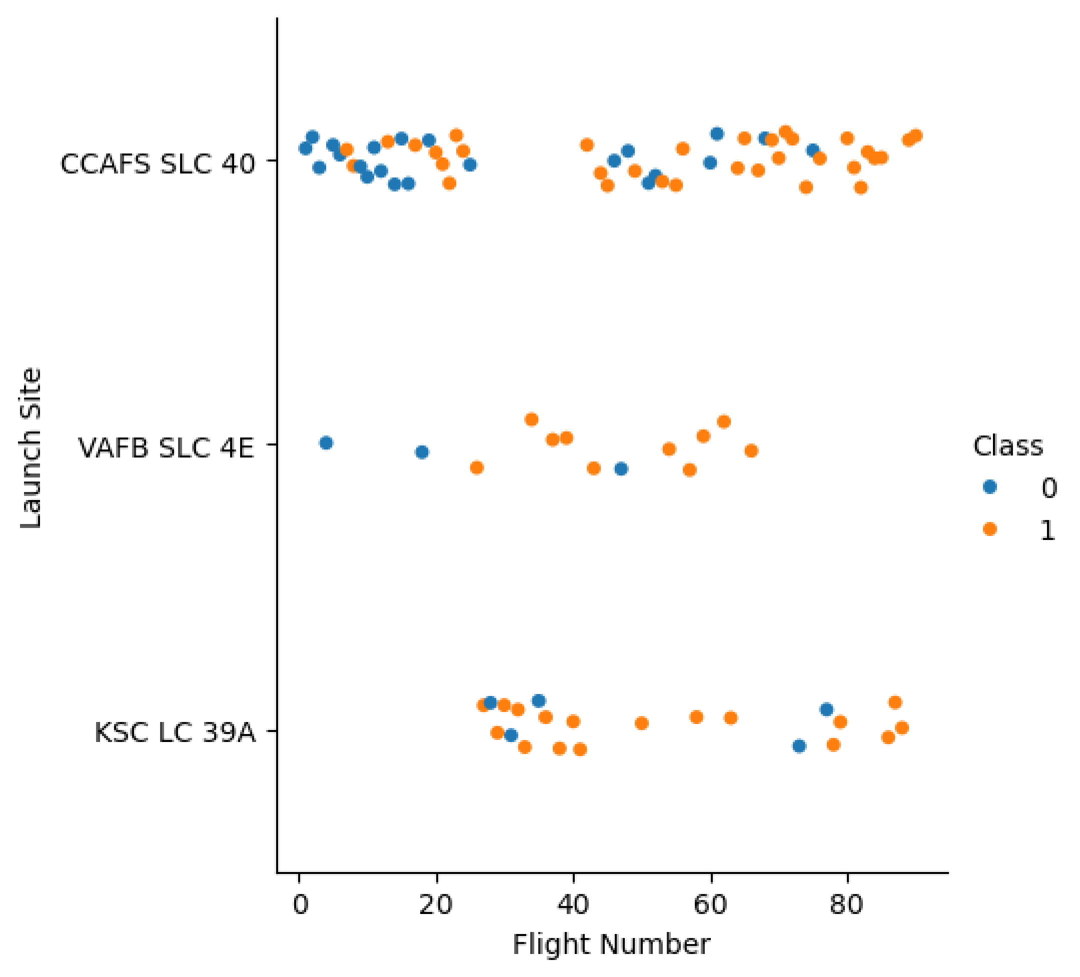
- Decision Tree Classifier had the best accuracy
- Model can accurately determine if the mission will fail or succeed

Section 2

Insights drawn from EDA

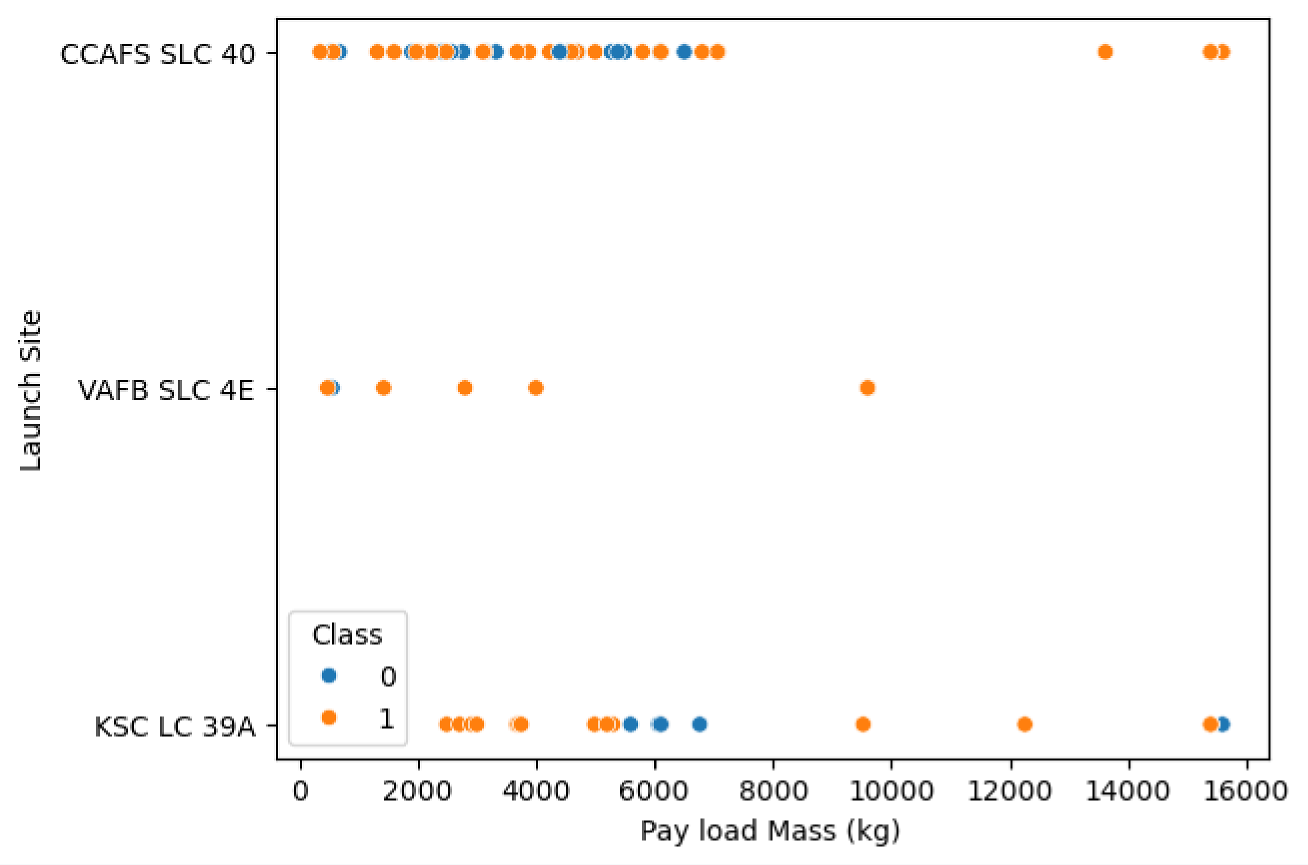
Flight Number vs. Launch Site

- Sites CCAFS SLC 40 and VAFB SLC 4E have been used for early launches while site KSC LC 39A appears to be a newer site, only starting to have launches after flight number 20.
- Site CCAFS SLC 40 had the greatest number of launches while site VAFB SLC 4E has the lowest number.



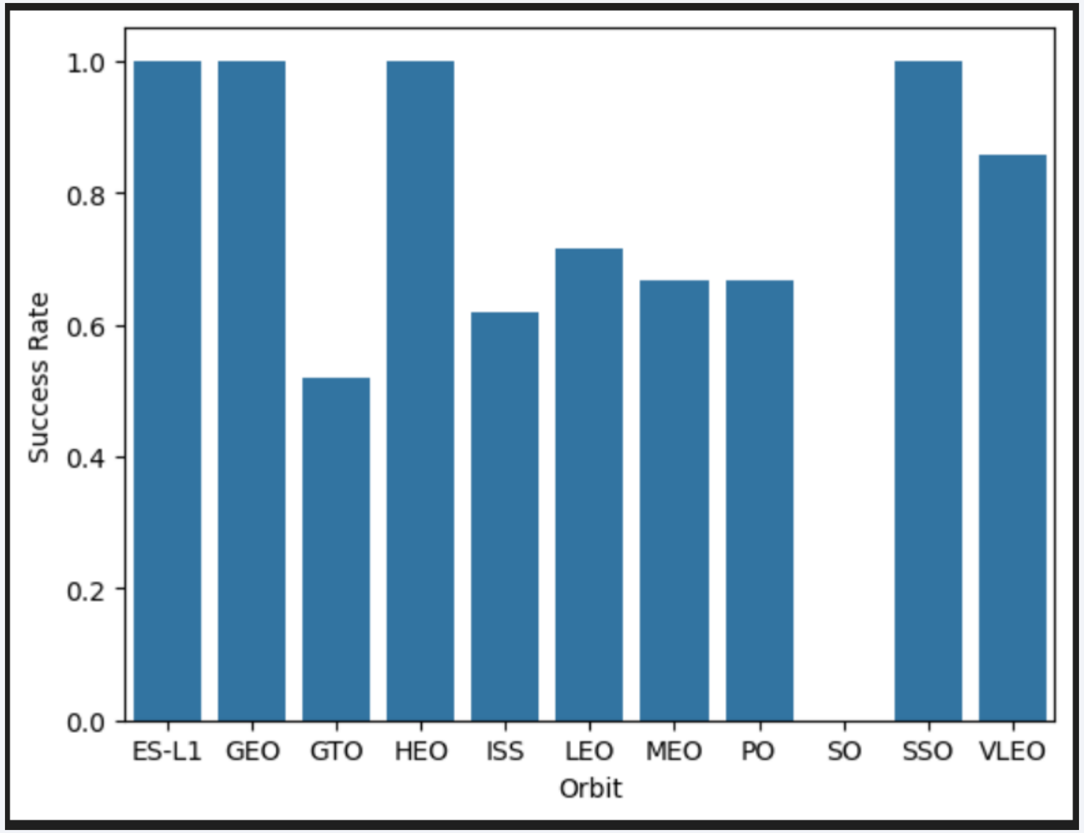
Payload vs. Launch Site

- Sites CCAFS SLC 40 and KSC LC 39A have more launches compared to site VAFB SLC 4E
- The two sites also perform missions with large payloads, while site VAFB SLC 4E only ever handled payloads of less than 10000 kg



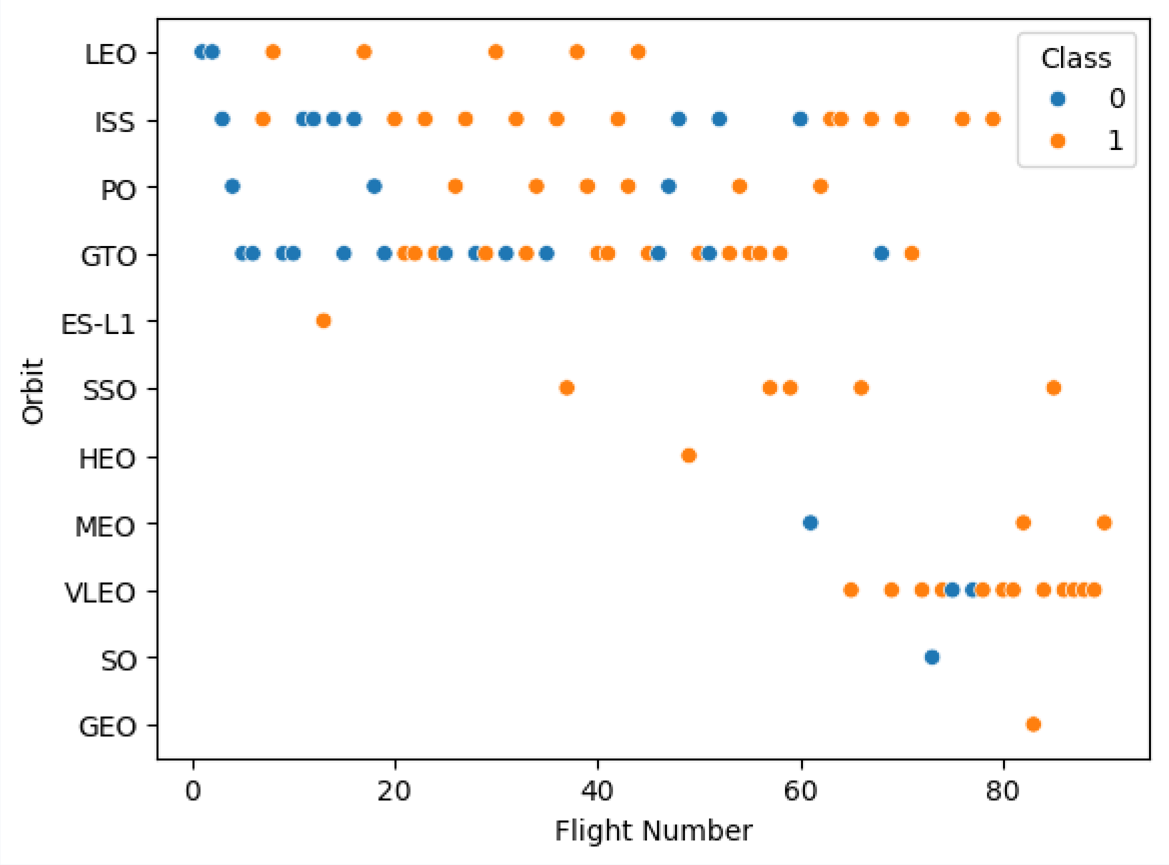
Success Rate vs. Orbit Type

- For orbits that are not commonly launched for, the success rates are either very high like with ES-L1, GEO, HEO, SSO or very low as with SO. MEO also has low samples and therefore with these orbits we can't extract accurate information without more launches.
- For other more common launches, like GTO and ISS, they have success rates of 50% and above, showing that SpaceX typically have positive outcomes with their missions



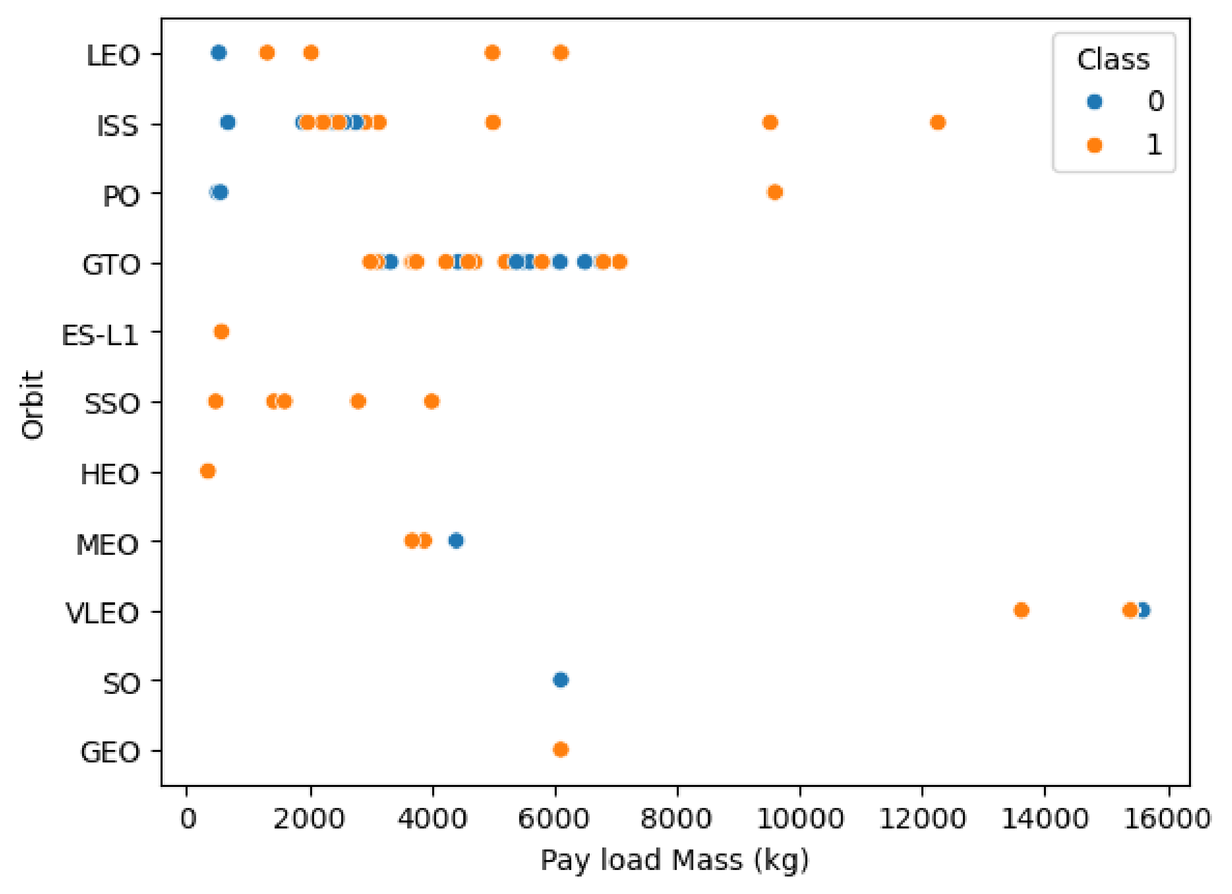
Flight Number vs. Orbit Type

- It can be seen for LEO orbits, later flights improved successful outcomes.
- For others like ISS, GTO, PO and VLEO, there appears to be no apparent relationship between later flights and outcomes.



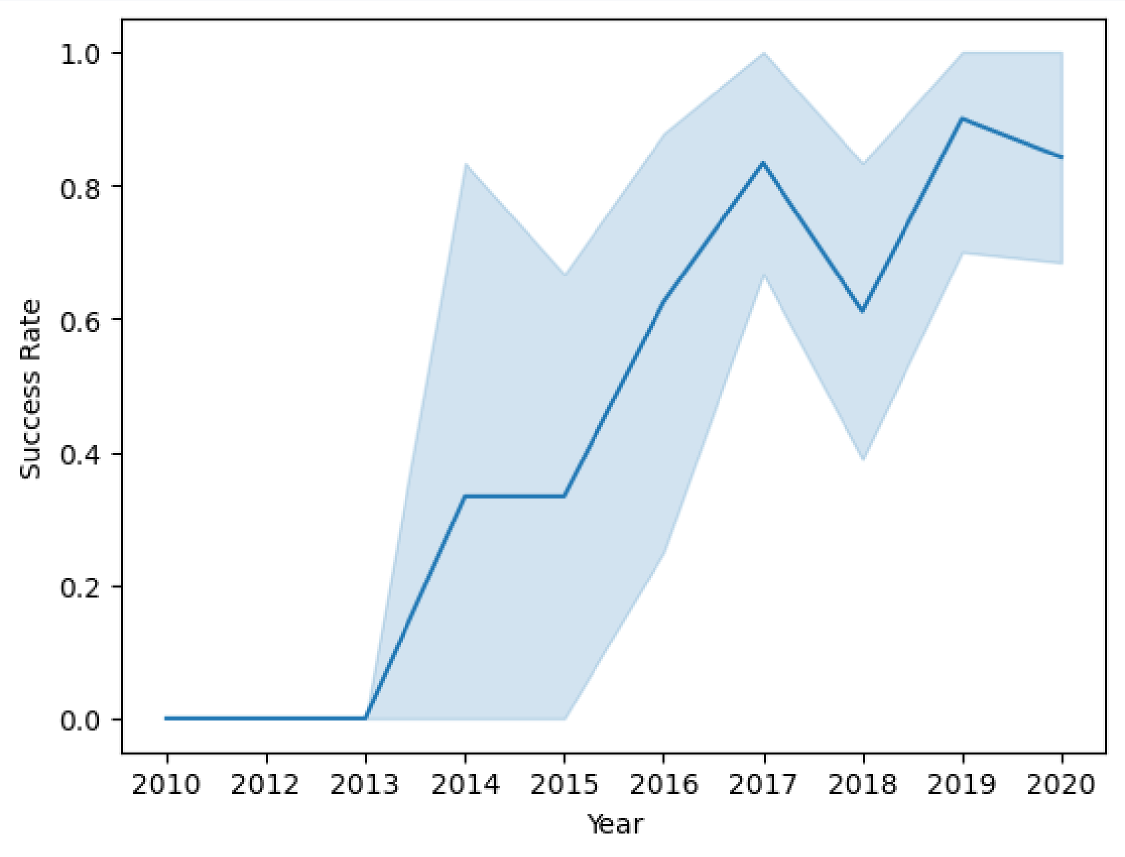
Payload vs. Orbit Type

- For Orbit types Polar, LEO and ISS, as payload goes up, their successful outcomes also goes up.
- However, for Orbit GTO this kind of relation is harder to identify as the payload does not seem to determine successful and failure outcomes.



Launch Success Yearly Trend

- Since the Year 2013, the number of successful launches increases over the years, where it had dips in the year 2018 and 2020.



All Launch Site Names

- There are 4 Launch Site used by SpaceX.
- This is important information to based our later findings on.

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

Launch Site Names Begin with 'CCA'

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

- They were all launched from Site CCAFS LC-40.
- Landing them were either failures or no attempts were made to do so
- They had all used F9 v1.0 boosters.
- This gives insight on the first couple launches done as the data is ordered by date with an ascending order

Total Payload Mass

- The total amount of payload mass carried by NASA is a whopping 45596 kg.
- This shows that NASA has lifted a lot of things from Earth into space.

Total Payload Mass (kg)
45596

Average Payload Mass by F9 v1.1

- The average payload mass for the booster F9 v1.1 is around 2534.66 kg.
- This shows that this booster is used to handle payloads on the lighter side.

Average Payload Mass (kg)
2534.6666666666665

First Successful Ground Landing Date

- This shows that SpaceX's first ever successful ground pad landing was on 2015-12-22.

Date of First Successful Landing Outcome
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster Versions of F9 FT are the only ones that made successful drone ship landings.
- This shows that for medium amounts of payload, booster F9 FT is the most consistent to successfully land on drone ships.

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- There has been a total of 71 successful and failure mission outcomes by SpaceX
- This shows SpaceX commitment to testing their technology in reaching their goals of Space Travel and making it cheaper.

Total Successful and Failure Mission Outcomes
71

Boosters Carried Maximum Payload

- The F9 B5 boosters are the ones that carry the maximum payload.
- This may mean that this booster version is used to test up till the maximum payload while other booster versions are used for other purposes.

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

2015 Launch Records

- 2 failure landings to drone ships occurred in the Year 2015.
- They failed in the months January and April where they with the same booster version “F9 v1.1” and on the same Launch Site “CCAFS LC-40”.
- This showed that they attempted to land similar ships but failed both attempts, which may show the need for larger changes to be made

Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- Between the dates 2010-06-04 and 2017-03-20, most landings were attempted as 21 landings were attempted and 10 were not attempted
- In terms of successful and failed landings, the most common were drone ship landings with an equal number of successful and failed landings.
- This may suggest that drone ships landings are the most favorable landing location for the team at SpaceX.

Landing_Outcome	Count
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

Section 3

Launch Sites Proximities Analysis

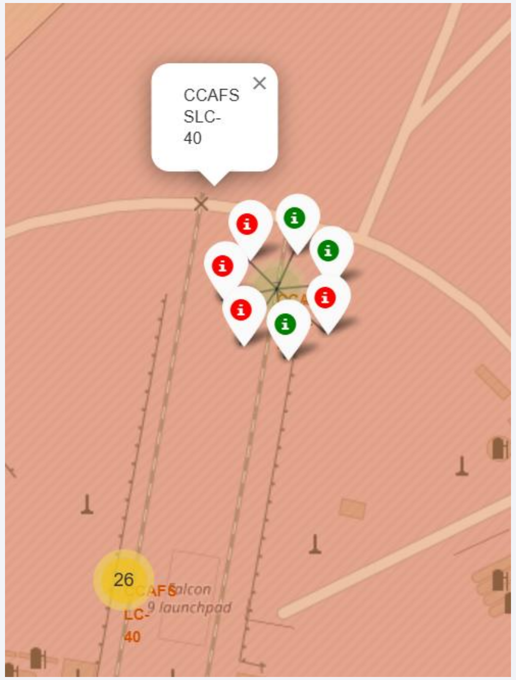
Launch Site Locations

There are 4 Launch Sites which includes:

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

- Site VAFB SLC-4E is near the west coast of the USA, while the other 3 launch sites are placed close to each other at the east coast of the USA in Florida.
- Sites CCAFS LC-40 and CCAFS SLC-40 are right next to each other.

Color Labeled Launch Outcomes



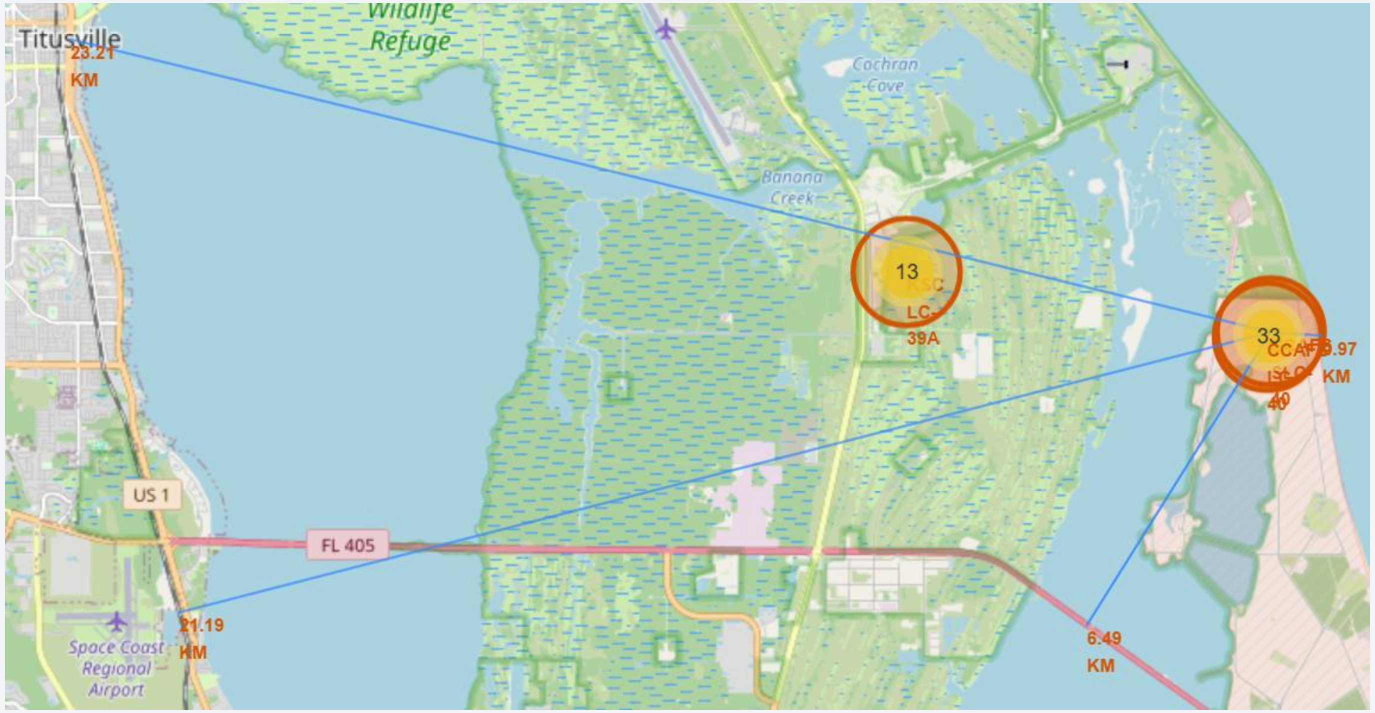
- The launch outcomes were plotted on the maps with green representing success and red meaning failure.
- It was seen that for each site, failures and successes occurred and that there is are higher number of launches from the east coast. This makes sense as there are 3 launching sites there while only one on the west coast.

Proximities of Site CCAFS SLC-40

Proximities and Distances:

- Coastline: 0.97 km
- Highway: 6.49 km
- Railway: 21.19 km
- Nearest City/Town: 23.21 km (Titusville)

This shows that there are considerable distances between launch sites and areas of human commutes. This makes sense since launch and landings are typically extremely loud and hot hence having the sites be so far away from where people travel/live is a good safety measure.

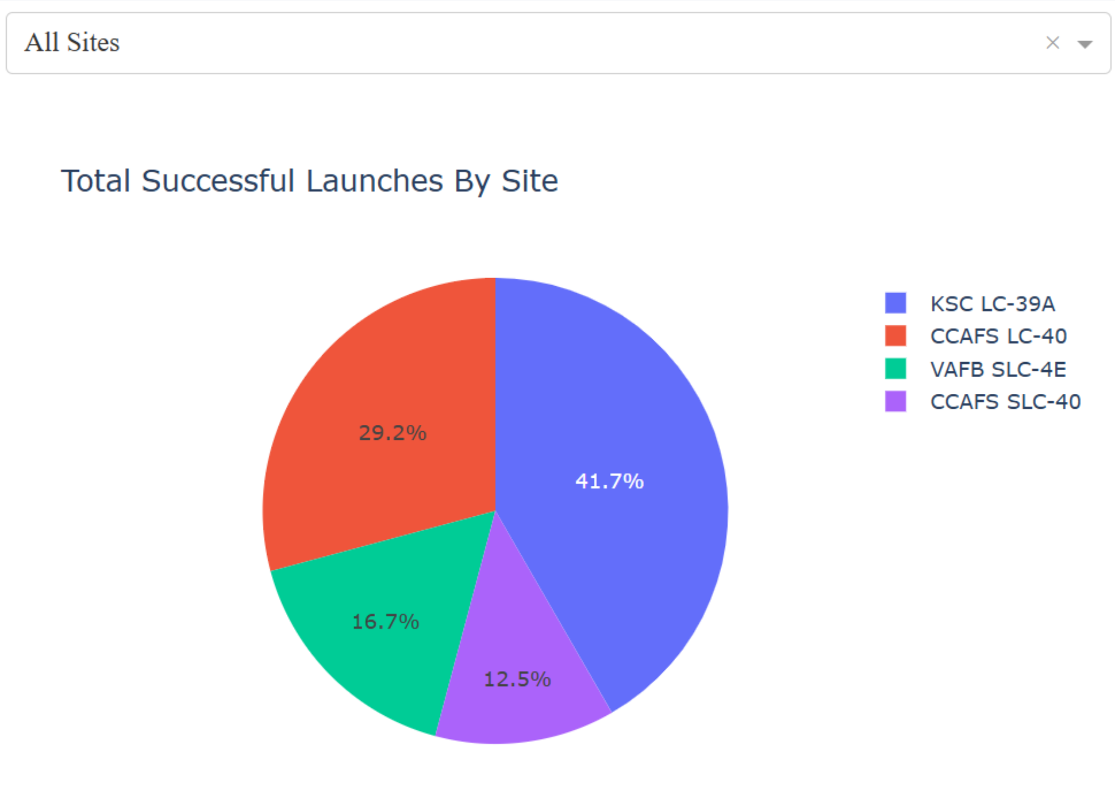


Section 4

Build a Dashboard with Plotly Dash

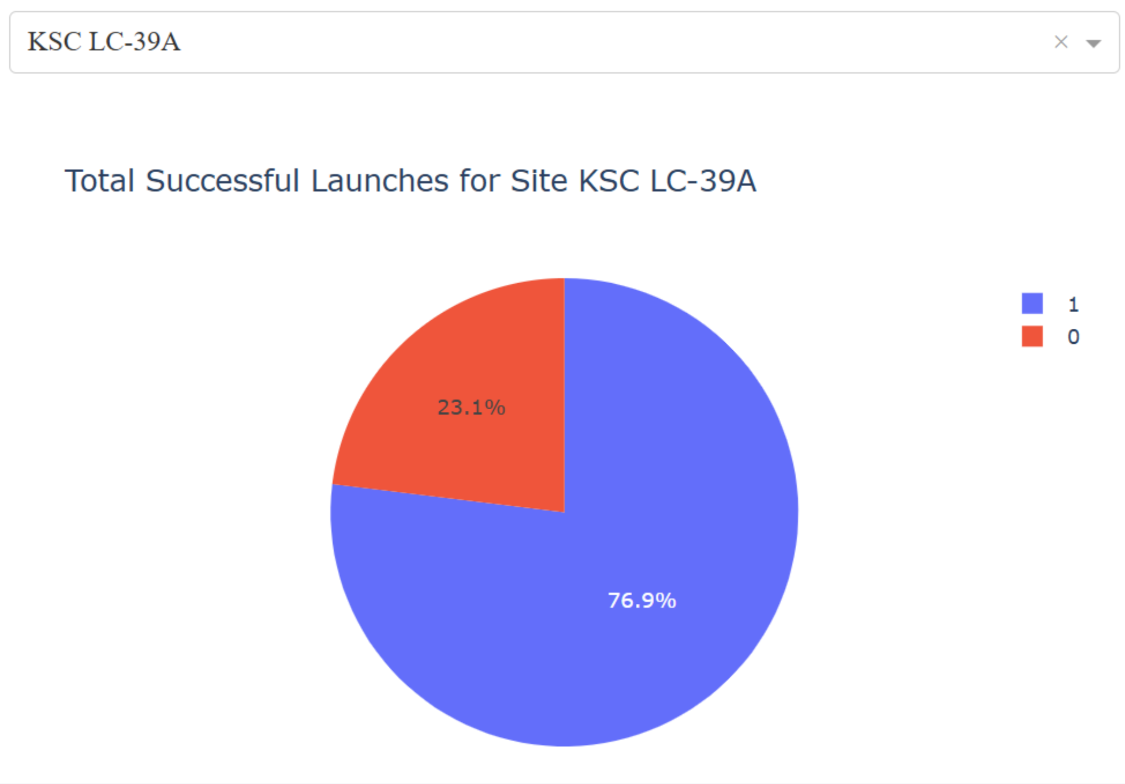
Total Successful Launches of each Site

- Site KSC LC-39A had the most successful launches compared to the other sites.
- Site CCAFS SLC-40 however, had the lowest number of successful launches.



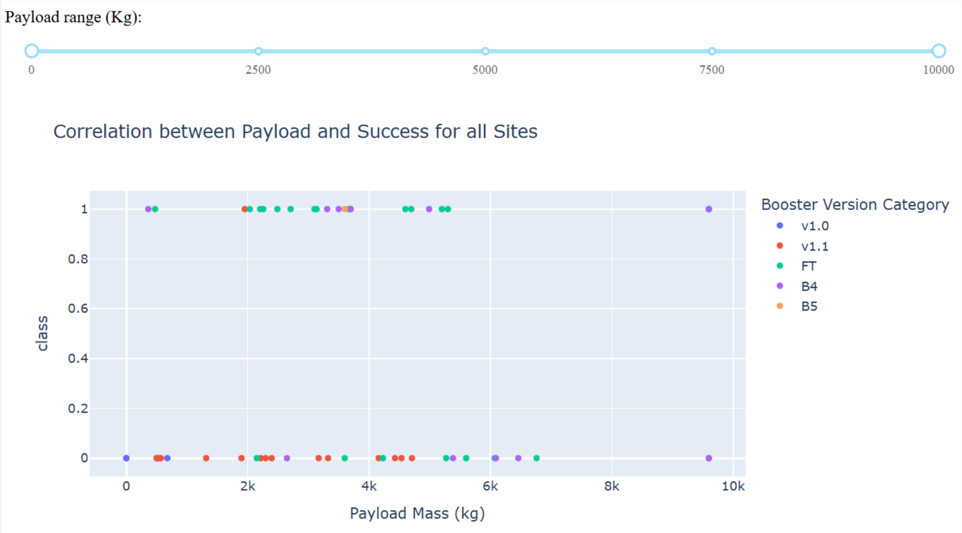
Highest Launch Success Ratio

- Site KSC LC-39A had the highest ratio of successful launches to failed launches out of the other Sites.
- A whopping 76.9% of the launches at this Site succeeded, showing that there might be something about the site that allows higher successes.



Payload vs Launch Outcomes

- It can be seen in the scatter plots that, most launches have payloads between 0 to 7000 kgs.
- FT boosters are more consistently having successful launches compared to other Booster Version Categories. It has the highest success rate.
- However, FT boosters are not as successful when the payload exceeds 5500 kgs.



Full Range

Range 2000 to 7500

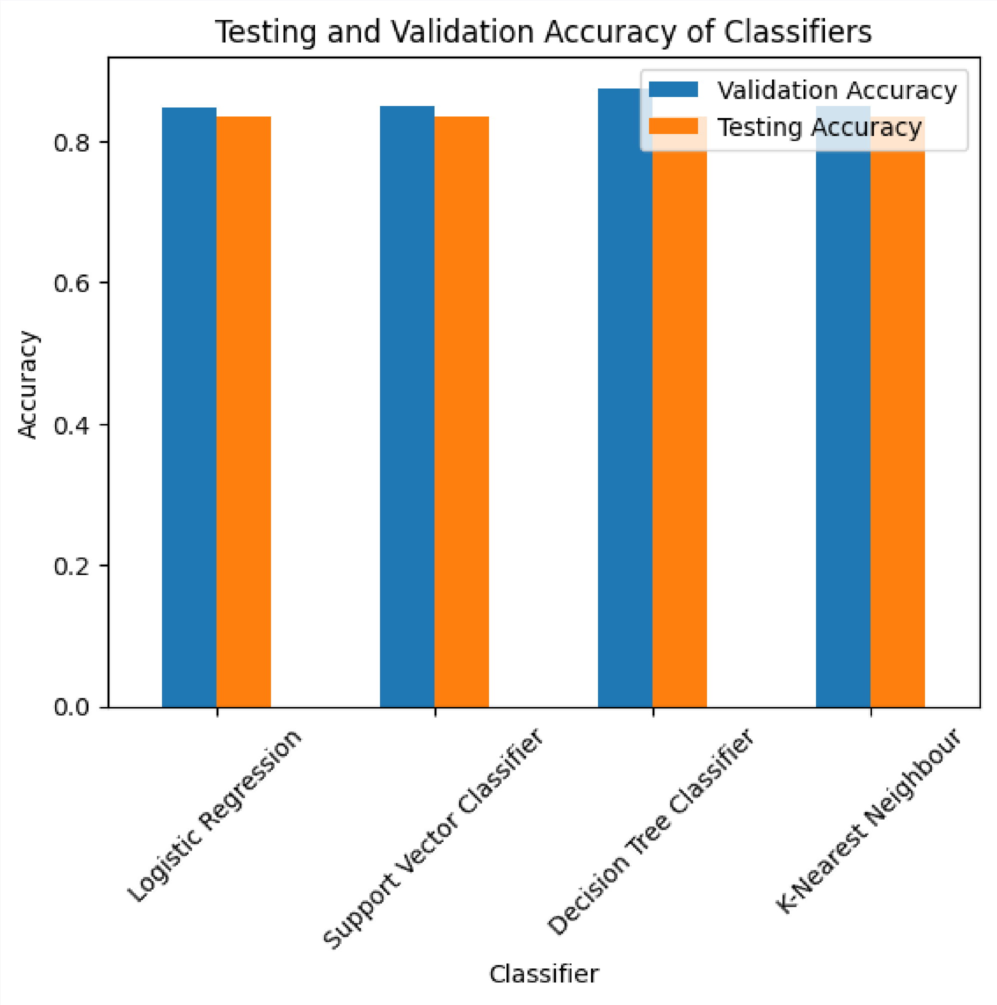
Section 5

Predictive Analysis (Classification)

Classification Accuracy

All models had gotten the same accuracy of 83.33% using the Test set. Their testing performance was also proven to be the same as their confusion matrixes were all the same.

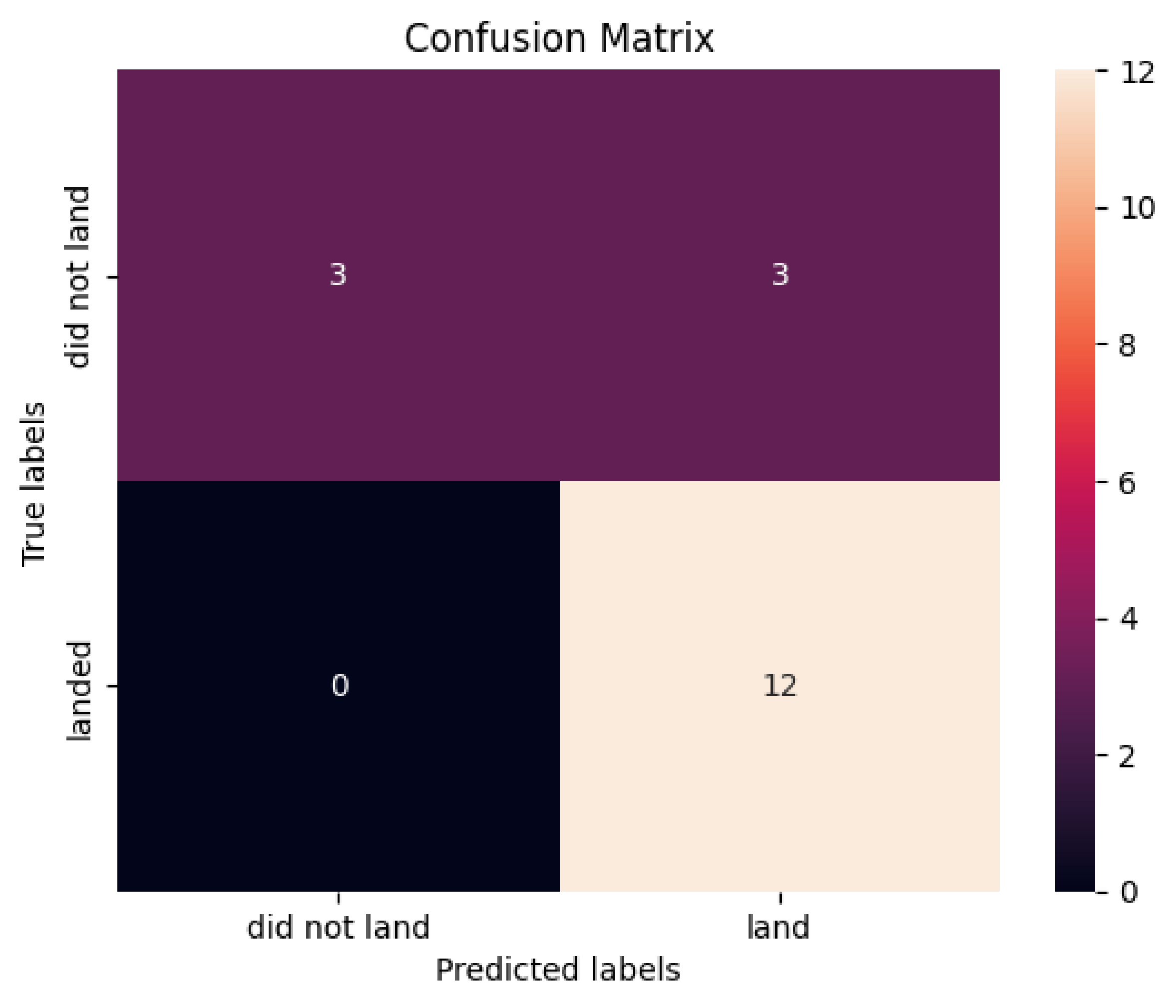
However, the Decision Tree Classifier had the highest Validation Accuracy showing that it was a cut above the rest.



Confusion Matrix

This was the confusion matrix of the Decision Tree Classifier (Best Classifier).

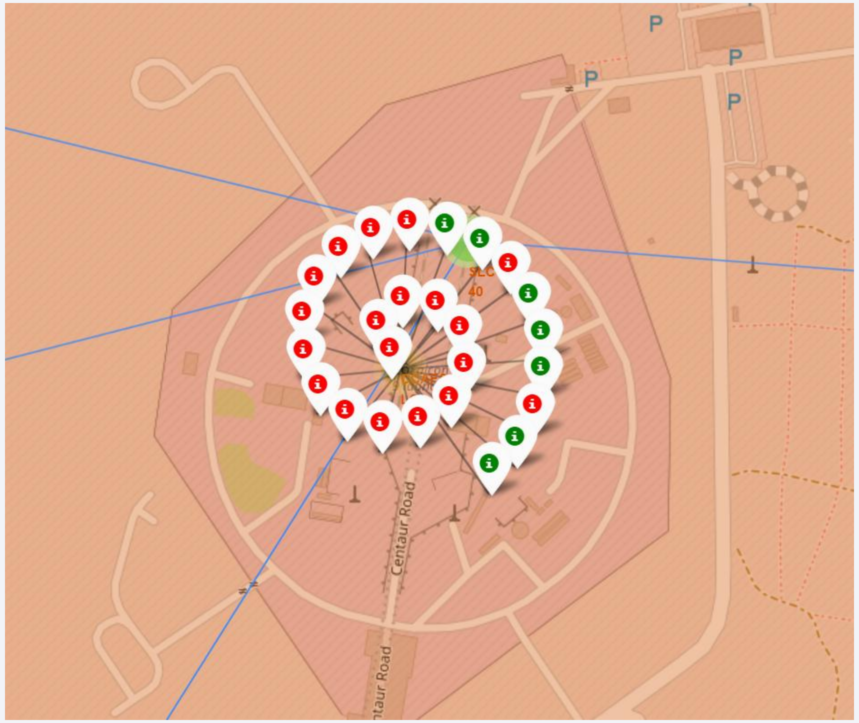
It was able to properly predict 15 labels and no False Negatives. However, it had predicted 3 False Positives showing that it has low Precision.



Conclusions

- The chosen Decision Tree Classifier can consistently predict the outcome of SpaceX launches with somewhat low precision. Using this classifier, users can implement its predictive capabilities to support or be against SpaceX with their services.
- As an outcome of this project, Interactive Maps and Dashboards were made to help is understanding and analyzing the SpaceX mission launch data users. Users can make use of them to more interactively learn of the data through a fun user experience.

Appendix



Markers of the launches done on Site
CCAFS LC-40

Green: Successful, Red: Failure

Thank you!