IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Otso Koskelo
09.10.2025

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

Collection of data from two sources (SpaceX API and Wikipedia) were used to gather information about SpaceX mission launches. This data was investigated and explored for further insight. Important features were then selected and used for machine learning models in order to predict whether a launch with specific parameters would lead to a successful landing or not. The results were then critically evaluated.

Introduction

Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against space X for a rocket launch.

This projects aims to source data from publicly available sources and use this to predict if a mission would lead to a successful landing or not. For this we are employing machine learning techniques.

Section 1

Methodology

Methodology

Summary

Data was first collected from two sources, SpaceX API and public Wikipedia site. Data was then processed, null values taken care of and mission outcomes were identified and labeled as 1 and 0. Data was further explored using visualization and SQL as well as interactive methods, namely Folium and Plotly Dash. Finally predictive analysis was performed with different machine learning algorithms in order to predict the outcome of the different missions.

The work was carried out using Python and Jupyter Notebook.

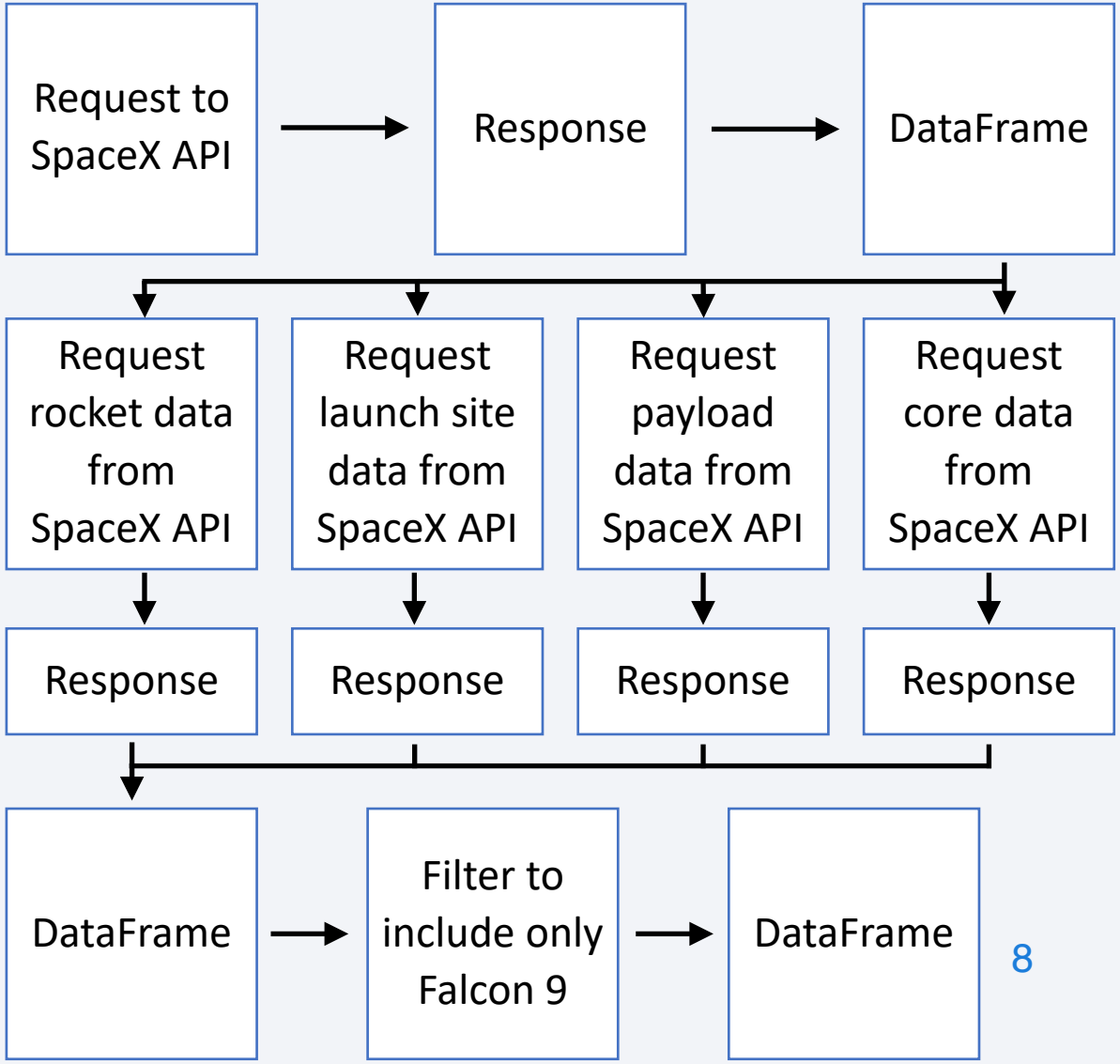
Data Collection

The rocket launch data set for the project was collected from two sources.

- Directly from SpaceX using API
- From a public (archived) Wikipedia page using web scraping

Data Collection – SpaceX API

The data was obtained by using a get()-command from the requests-library. The response object from the API was then assigned to a variable which was used to create a pandas DataFrame object. Helper functions were used to collect additional data which was then combined with the original DataFrame. In the last step, data was filtered to include only Falcon 9 flights.

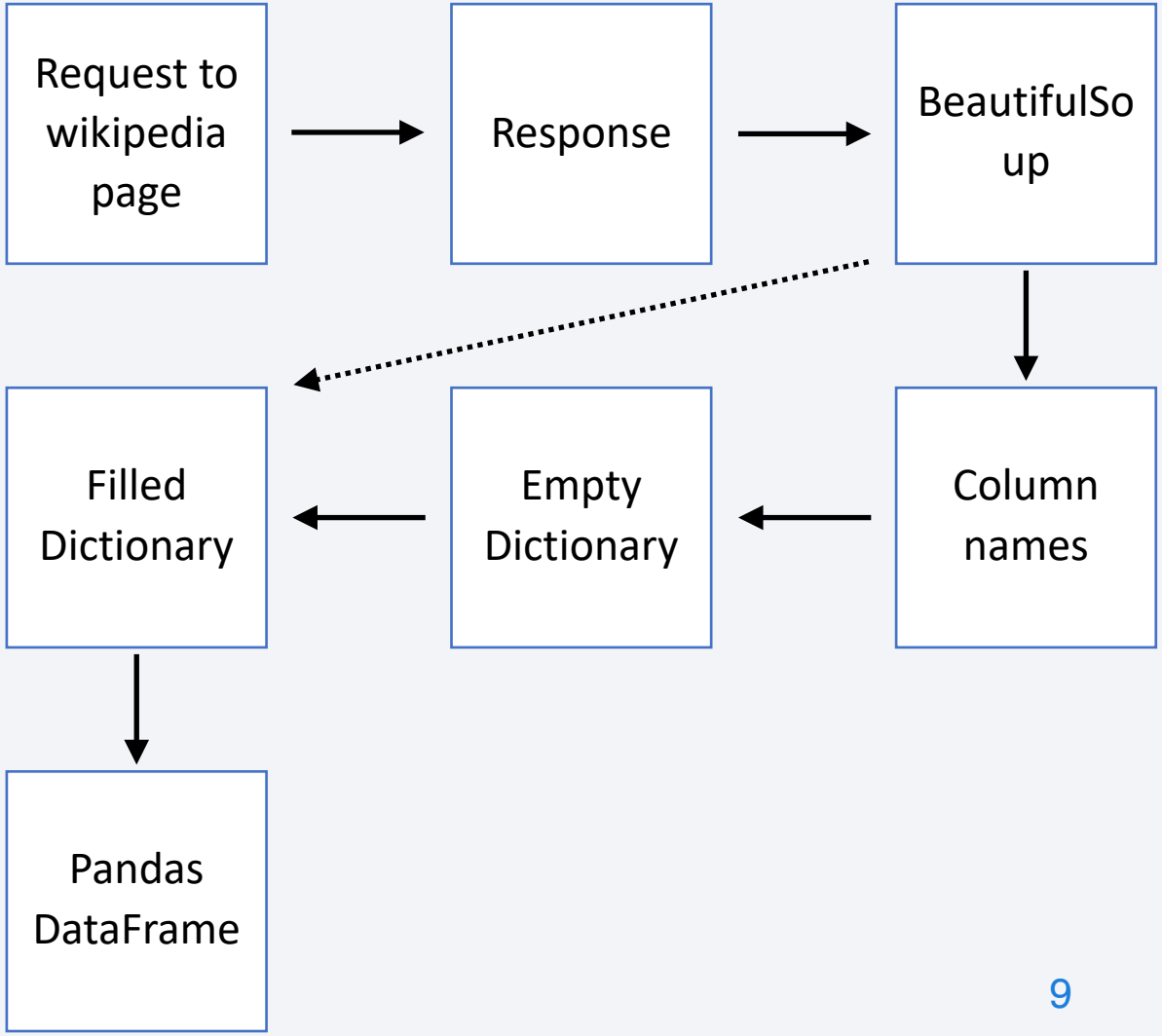


https://github.com/OrigoKO/Coursera/blob/b1b96e95fe018517e2bdd81087571a718f5c53fb/Capstone/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

Like before, the initial data from the wikipedia page was obtained by using a get()-command from the request library. The response-object was then used to create a BeautifulSoup-object.

This object was used to extract column names from a table which were stored in a dictionary-object. The actual data from the table was then added to the dictionary which was then transformed in to a pandas data frame.



https://github.com/OrigoKO/Coursera/blob/b1b96e95fe018517e2bdd81087571a718f5c53fb/Capstone/jupyter-labs-webscraping.ipynb

Data Wrangling

Data wrangling process was started by identifying the given features of the data. Null values of PayloadMass were replaced with an average value. The distribution between different launch sites, as well as occurrence of each orbit was noted.

Different outcome categories for the missions was identified and a new feature was engineer stating whether a mission outcome was a success or a failure (1 or 0, respectfully).

https://github.com/OrigoKO/Coursera/blob/b1b96e95fe018517e2bdd81087571a718f5c53fb/Capstone/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

Data visualization was used for exploratory data analysis. Following relationships were plotted using scatter plots, bar plots and line plots. The goal was to see underlying correlation between variables and a mission success rate:

1) Visualize the relationship between Flight Number and Launch Site
2) Visualize the relationship between Payload Mass and Launch Site
3) Visualize the relationship between success rate of each orbit type
4) Visualize the relationship between FlightNumber and Orbit type
5) Visualize the relationship between Payload Mass and Orbit type
6) Visualize the launch success yearly trend

Two distinct colors were used in plotting in order to differentiate successful missions from unsuccessfuls.

https://github.com/OrigoKO/Coursera/blob/b1b96e95fe018517e2bdd81087571a718f5c53fb/Capstone/edadataviz.ipynb

EDA with SQL

The following SQL-queries were performed to the data:

https://github.com/OrigoKO/Coursera/blob/b1b96e95fe018517e2bdd81087571a718f5c53fb/Capstone/jupyter-labs-eda-sql-coursera_sqllite.ipynb

1) Displaying the unique names of the launch sites
2) Displaying the 5 sites where the name begins with ”CCA”
3) Displaying the total payload mass carried by boosters launched by NASA.
4) Displaying average payload mass carried by booster version F9 v1.1
5) Listing the date when the first successful landing outcome in ground pad was achieved
6) Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7) Listing the total number of successful and failure mission outcomes
8) Listing all the booster_versions that have carried the maximum payload mass, using a subquery with a suitable aggregate function.
9) Listing the records which will display the month names, failure landing_outcomes in drone ship, booster versions, launch_site for the months in year 2015
10) Ranking the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

Build an Interactive Map with Folium

Folium was used to build an interactive map to gain insight of the launch sites. Markers and circles were drawn on a map to point out the locations of the launch sites.

Another set of markers were drawn on top of the launch sites with two different colors displaying the amount of successful and failed missions (green and red, respectively). This was done in order to easily compare the amount of successful and unsuccessful launches between launch sites.

Also, some lines were drawn to display distances between launch sites and selected points on interest such as coast line or a nearby major city.

https://github.com/OrigoKO/Coursera/blob/b1b96e95fe018517e2bdd81087571a718f5c53fb/Capstone/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

A dashboard was built with Plotly Dash to provide an interactive way to explore data easily. Two different graphic layouts were created:

1) A pie chart showing a ratio between successful and failed missions based on a selection of a launch site. If all sites were selected, the chart showed the relative amounts of successful missions between all sites
2) A scatter plot showing a relation between payload, mission outcome and rocket used (displayed with distinct colors). Also a slider was added for the user two restrict the results only to some range of payload.

https://github.com/OrigoKO/Coursera/blob/470d4a80d76de5f3d38c1bc99c7a8c524166d355/Capstone/spacex-dash-app.py

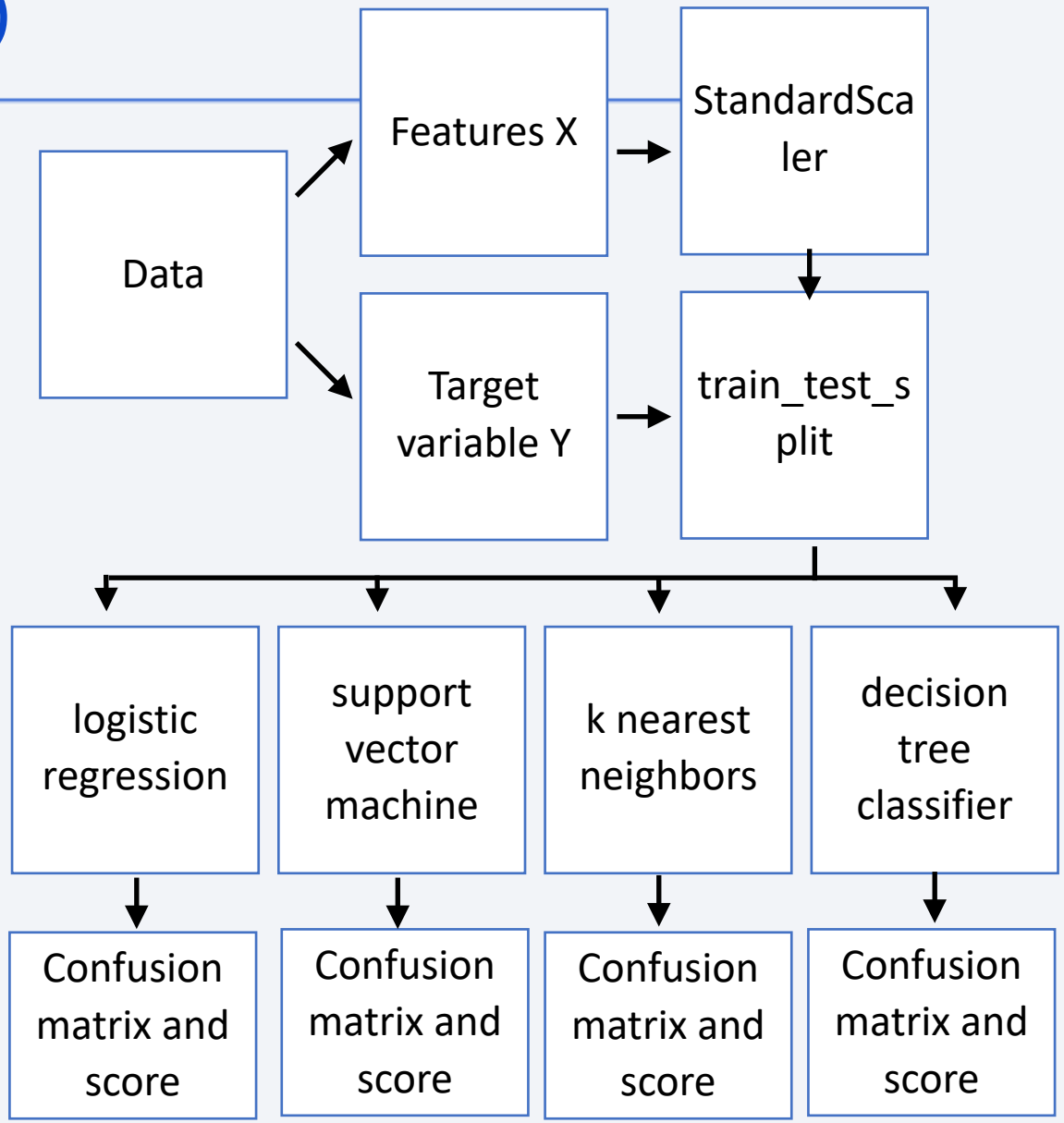
Predictive Analysis (Classification)

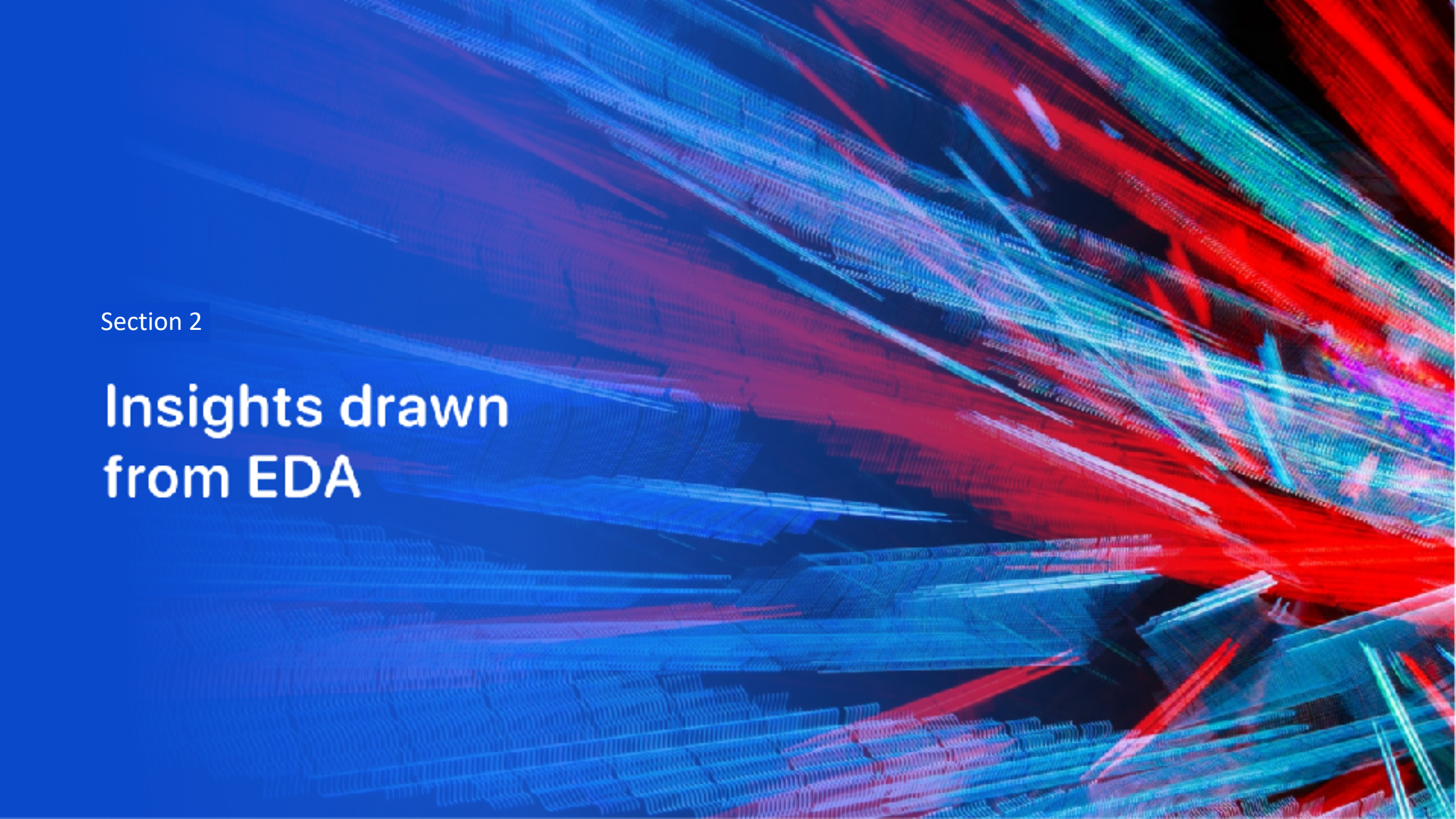
As a final step, machine learning algorithms were selected to predict an outcome of the future missions. The 'class' column was selected as predicted variable and the rest of the columns were used as input features.

The feature data was standardized with StandardScaler and train_test_split was used to split the data into train and test parts

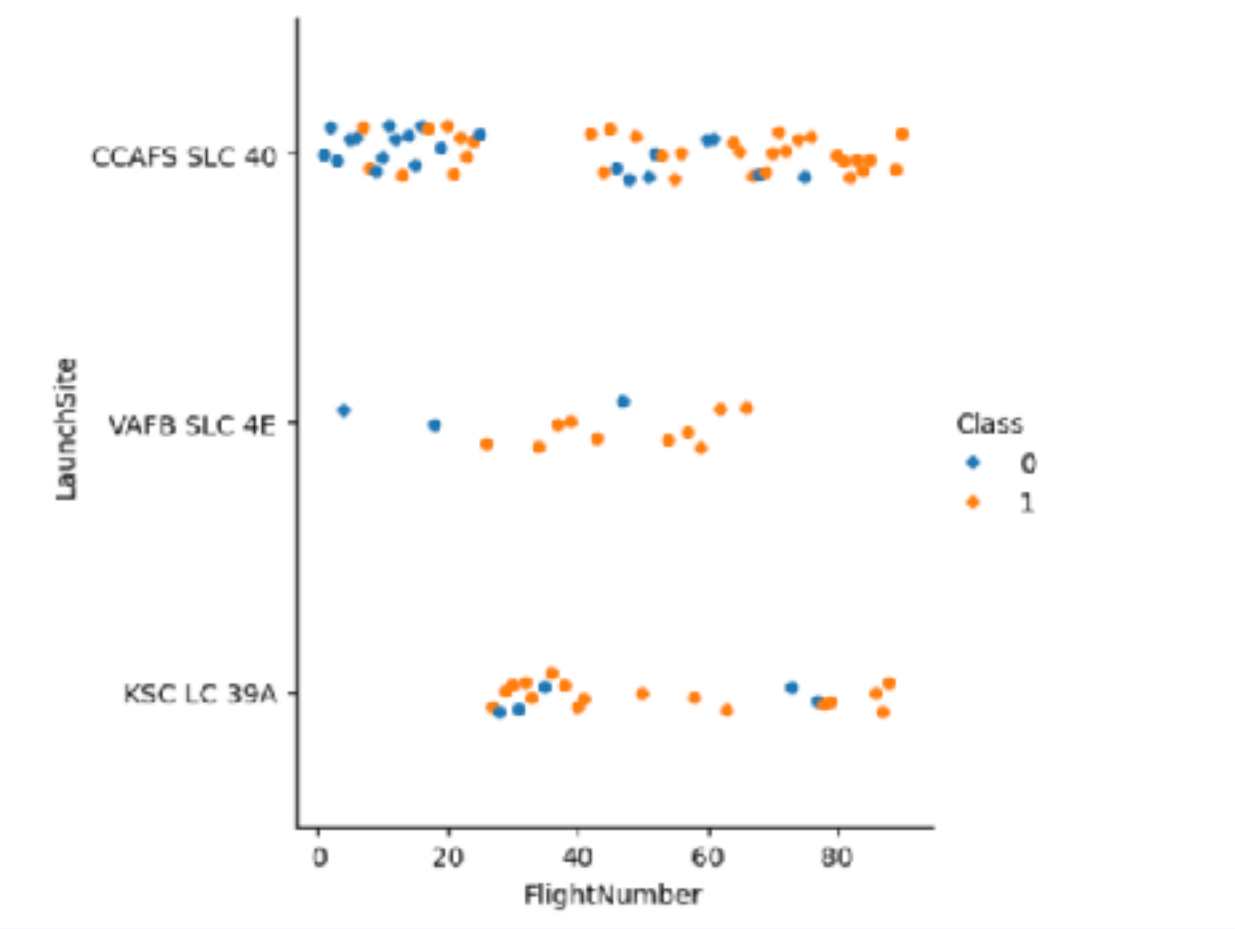
GridSearchCV was used for four different machine learning algorithms: logistic regression, support vector machine, decision tree classifier and k nearest neighbors. Confusion matrix as well accuracy scores were produced for each

https://github.com/OrigoKO/Coursera/blob/470d4a80d76de5f3d38c1bc99c7a8c524166d355/Capstone/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb




Section 2
Insights drawn
from EDA

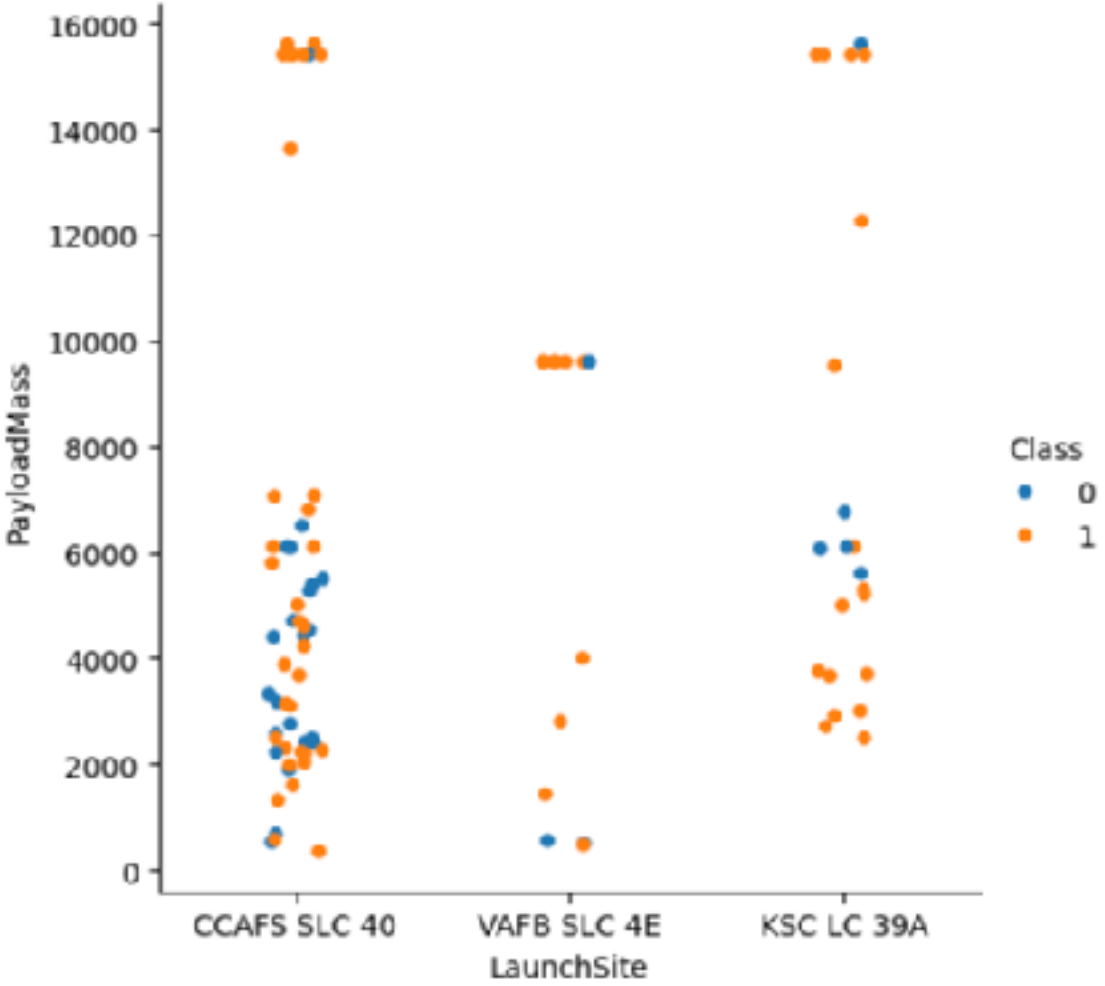
Flight Number vs. Launch Site



Overall one can see slight improvement of successful flights with higher flight numbers. This is especially clear in the topmost launch site (CCAFS SLC 40). This has had most missions over all as well.

VAFB SLC 4E has had the fewest amount of missions and KSC LC 39A joined the game a bit later having a first mission roughly around flight number 25.

Payload vs. Launch Site

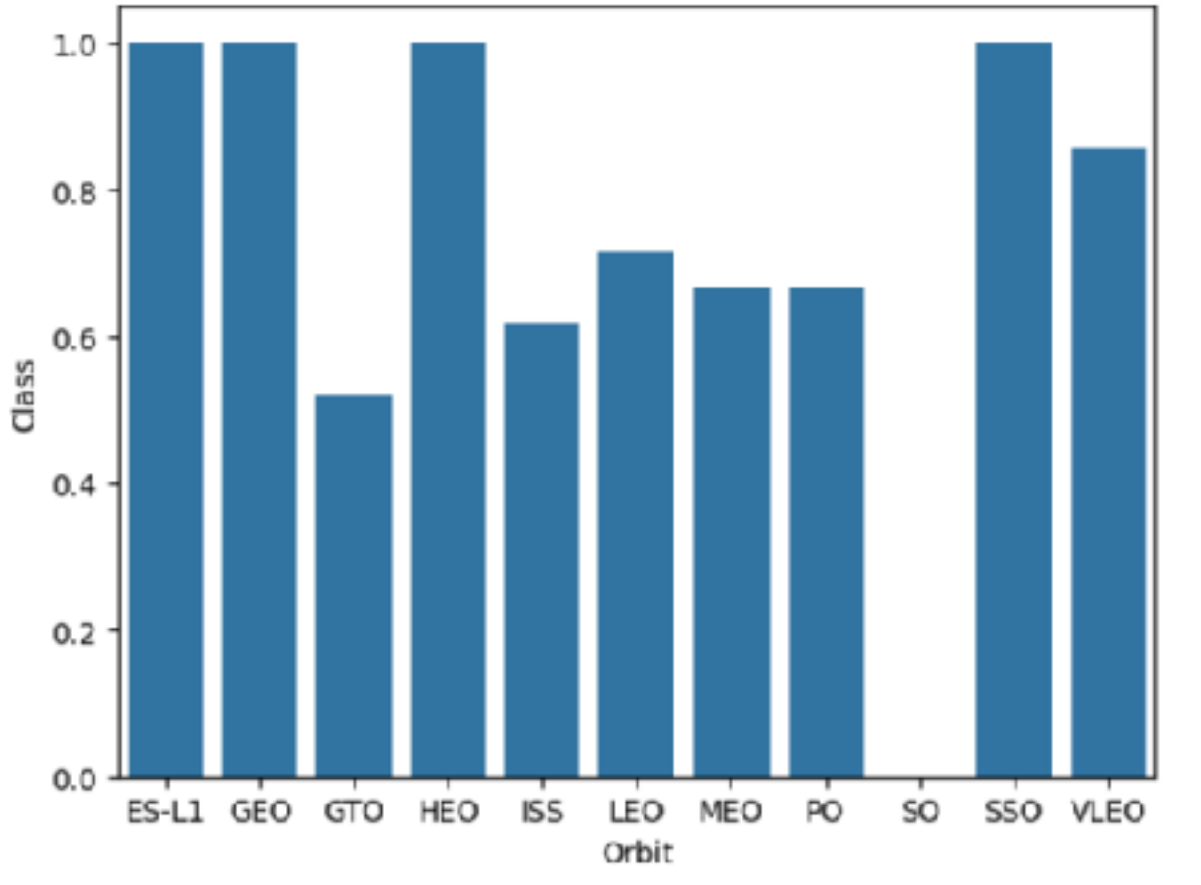


First thing to note is that launch site VAFB SLC 4E didn't have any flights with payloads greater than 10 000.

The launch site CCAFS SLC 40 has had more success with higher payloads.

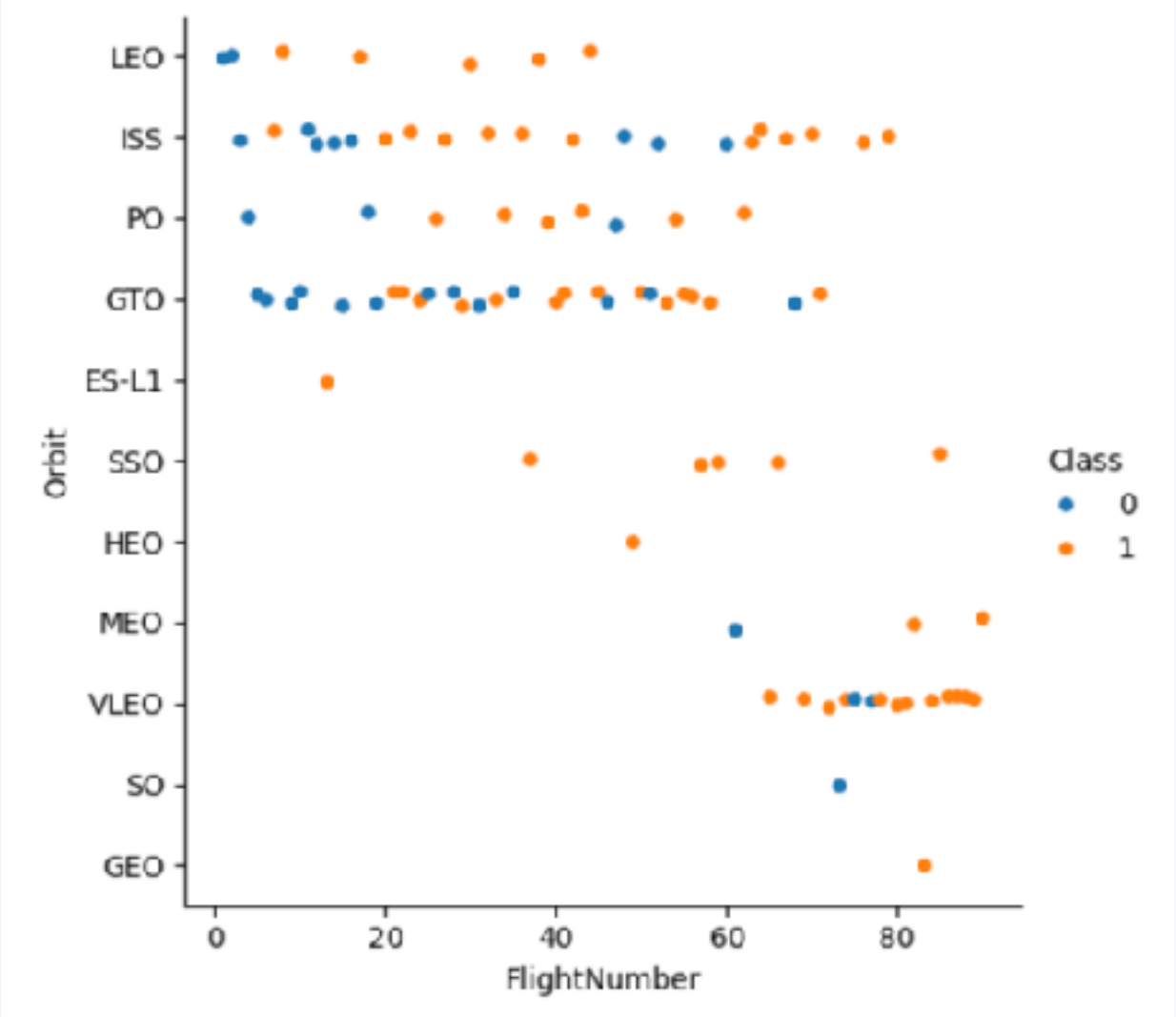
Most of the missions have been done with lighter payloads.

Success Rate vs. Orbit Type



From this chart one can easily see that the missions to orbits ES-L1, GEO, HEO and SSO have been most successful.

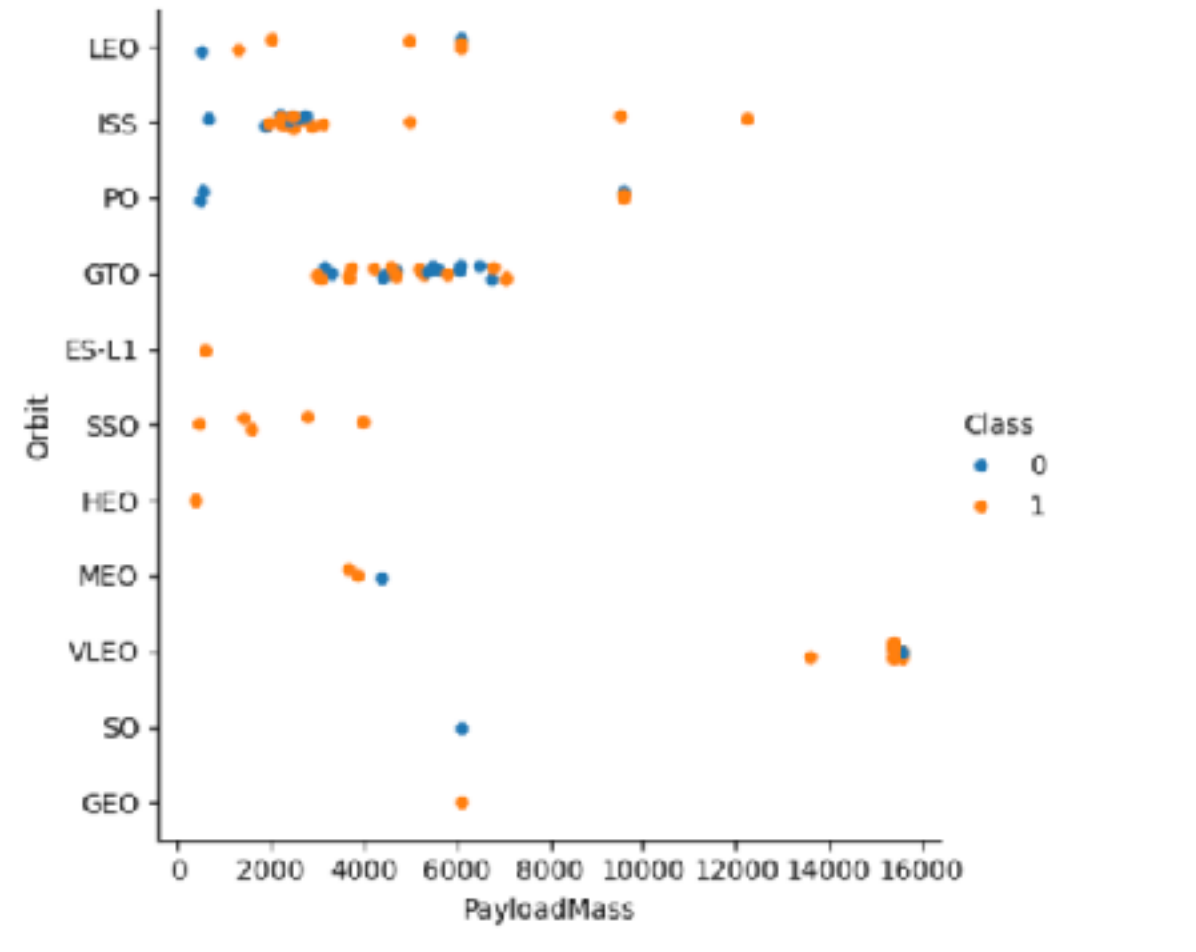
Flight Number vs. Orbit Type



When plotting flight number against the orbit type, we can see that some of the orbits have been used in the missions from the very start when some of the orbits have been used only for later missions.

We can also see, for example, that flights to orbit LEO have been successful after two failed attempts and flights to orbit SSO have all been successful. Other orbits have mostly more mixed result.

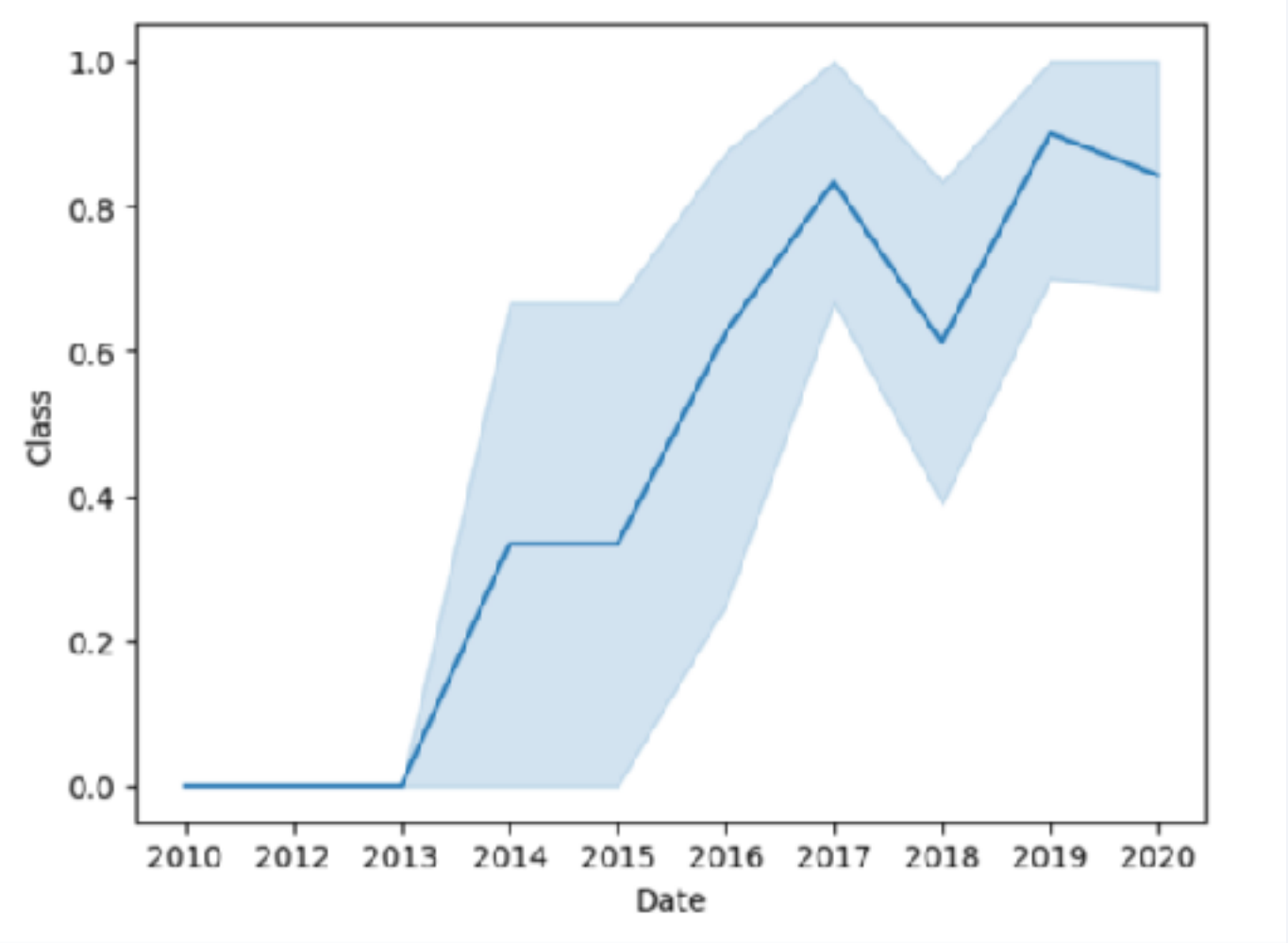
Payload vs. Orbit Type



From this graph we can see how the success of the missions is depending on the payload.

No very clear trends can be seen. Maybe the lowest payloads to orbits LEO, ISS and PO have also been first attempts and therefore unsuccessful? This can be somewhat confirmed from the earlier graph showing mission outcomes depending on the flight number.

Launch Success Yearly Trend



From this trend line we can see that the success rate of the missions have been rising in time as to be expected.

All Launch Site Names

The result of a SQL-query showing names of the different launch sites.

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

Launch Site Names Begin with 'CCA'

The result of a SQL-query showing all the records where the name of the launch site starts 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

Total Payload Mass

The result of a SQL-query showing total payload mass of all of the missions.

SUM(PAYLOAD_MASS__KG_)
99980

Average Payload Mass by F9 v1.1

The result of a SQL-query showing average payload mass of all the missions involving a booster version F9 v1.1

AVG(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

The result of a SQL-query showing the first date of a successful ground landing.

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

The result of a SQL-query showing all the boosters that have successfully landed on a drone ship with payload between 4000 and 6000.

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

Total Number of Successful and Failure Mission Outcomes

The results of a SQL-queries showing total number of successful and failed mission outcomes (respectfully).

COUNT(*)
100

COUNT(*)
1

Boosters Carried Maximum Payload

The result of a SQL-query showing the names of all the booster that have carried a maximum payload.

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

The result of a SQL-query showing all the failed missions from the year 2015 by month.

Month	Mission_Outcome	Booster_Version	Launch_Site
01	Success	F9 v1.1 B1012	CCAFS LC-40
04	Success	F9 v1.1 B1015	CCAFS LC-40

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

The result of a SQL-query showing the counts of all the landing outcomes between 04.06.2010 and 20.03.2017 in descending order.

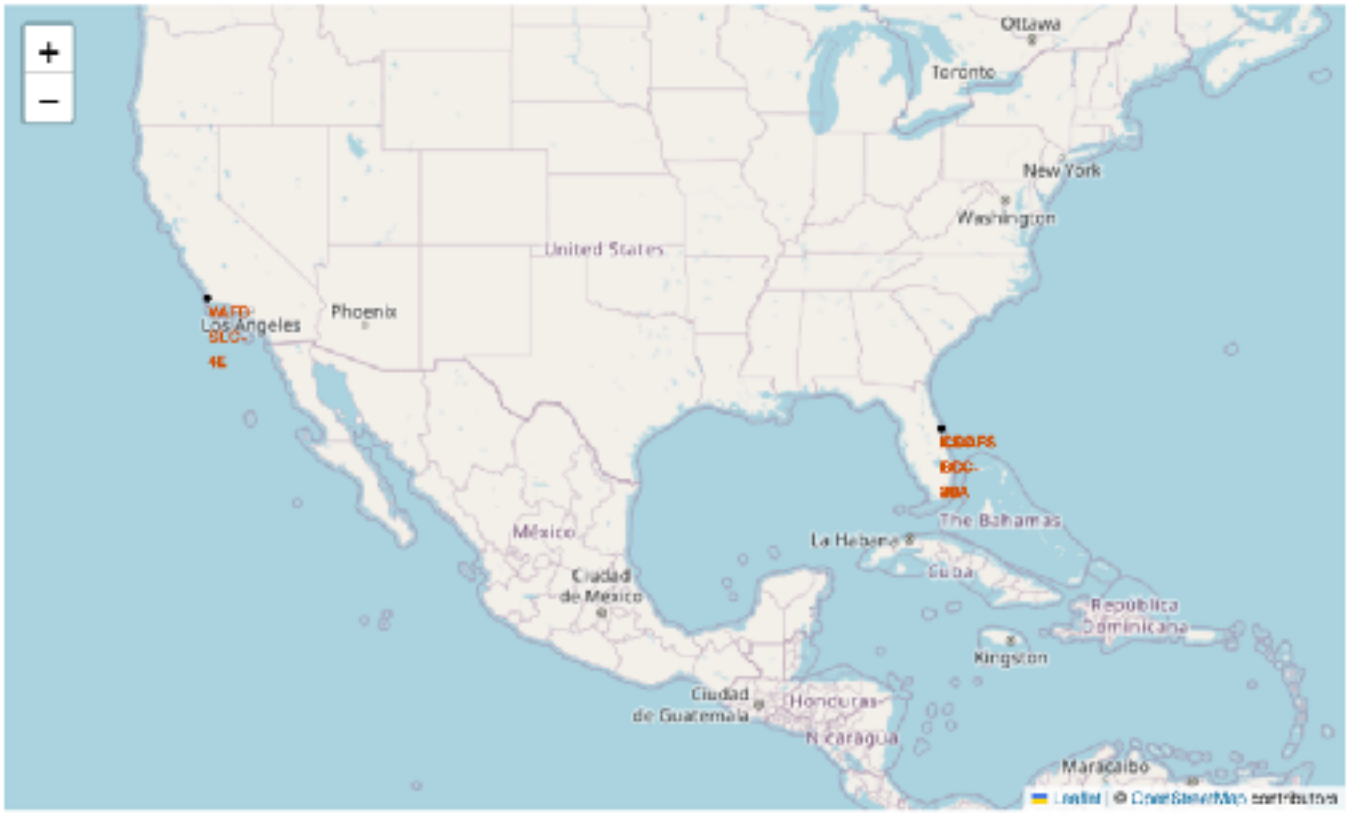
Landing_Outcome	Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

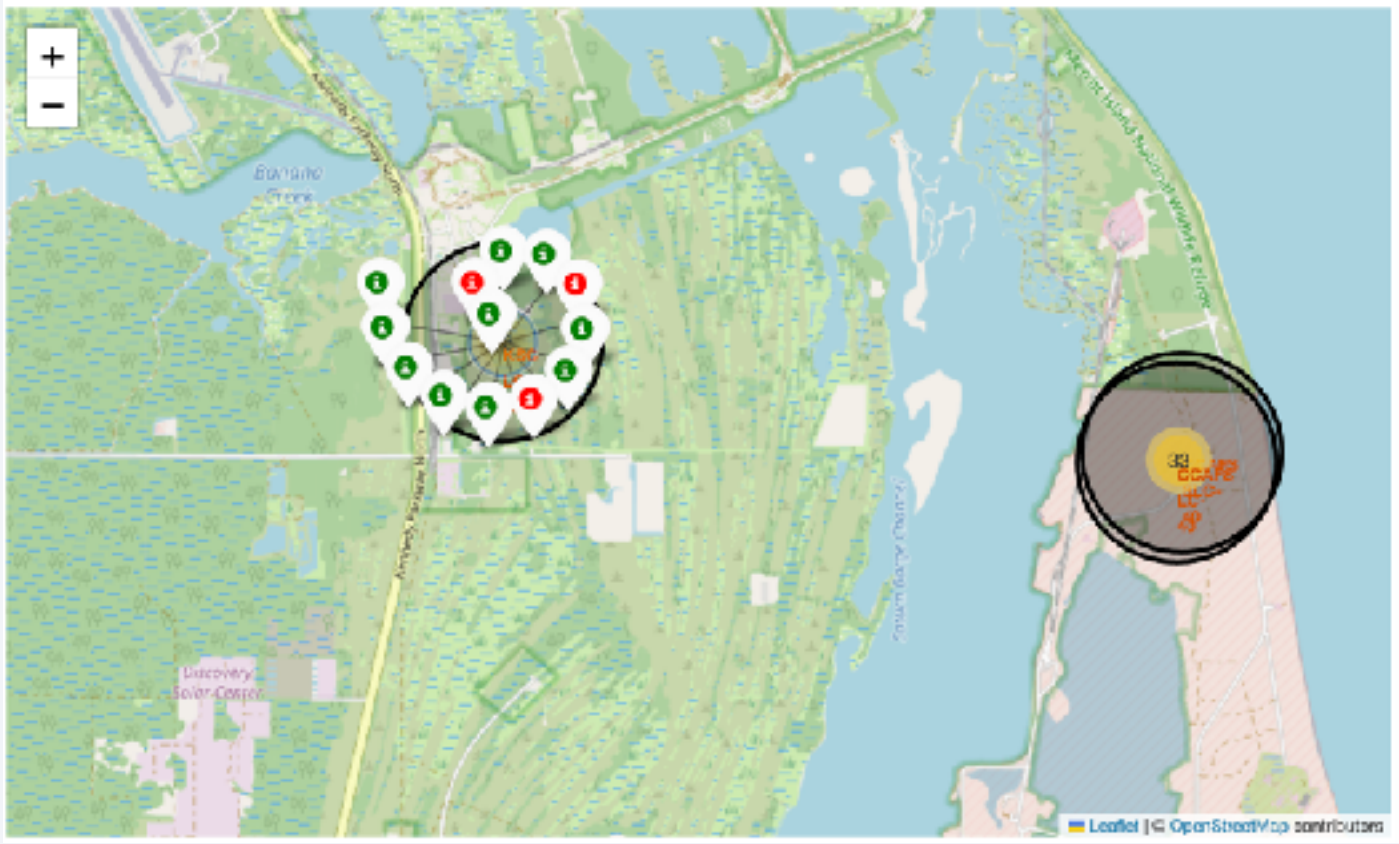
Launch site locations

Here is presented all the launch site locations on a Folium-map. We can see that the sites are divided in two locations on both sides of North American continent.



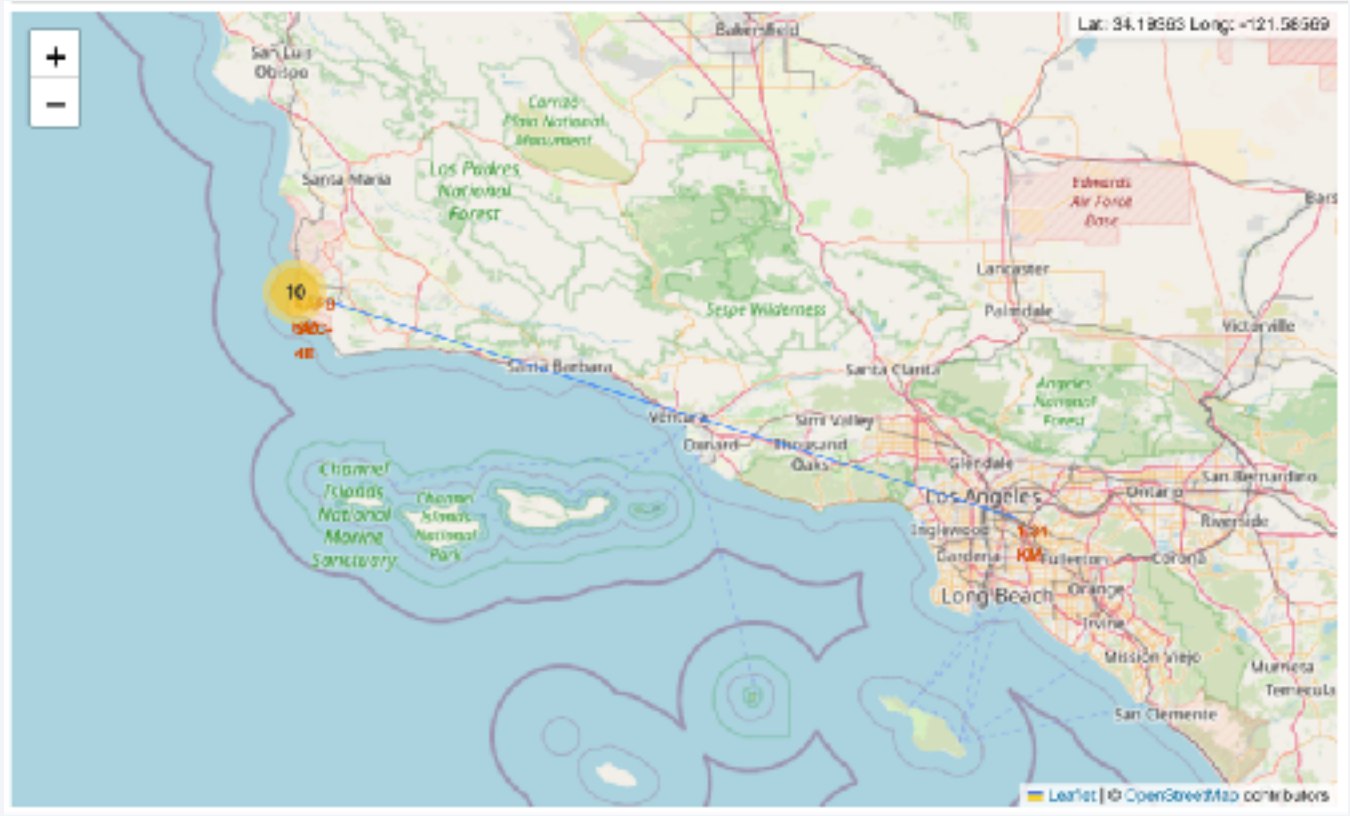
Color labeled mission outcomes for launch sites

Folium was used to present all mission outcomes per launch site. Two distinct colors were used to represent successful and failed mission launches (green and red, respectfully).



Distance between launch sites and selected elements

Folium was again used to calculate and present distances from launch sites to selected points of interest. Here is, for example, a line representing the distance between a launch site VAFB SLC-4E and a nearest major city, Los Angeles.

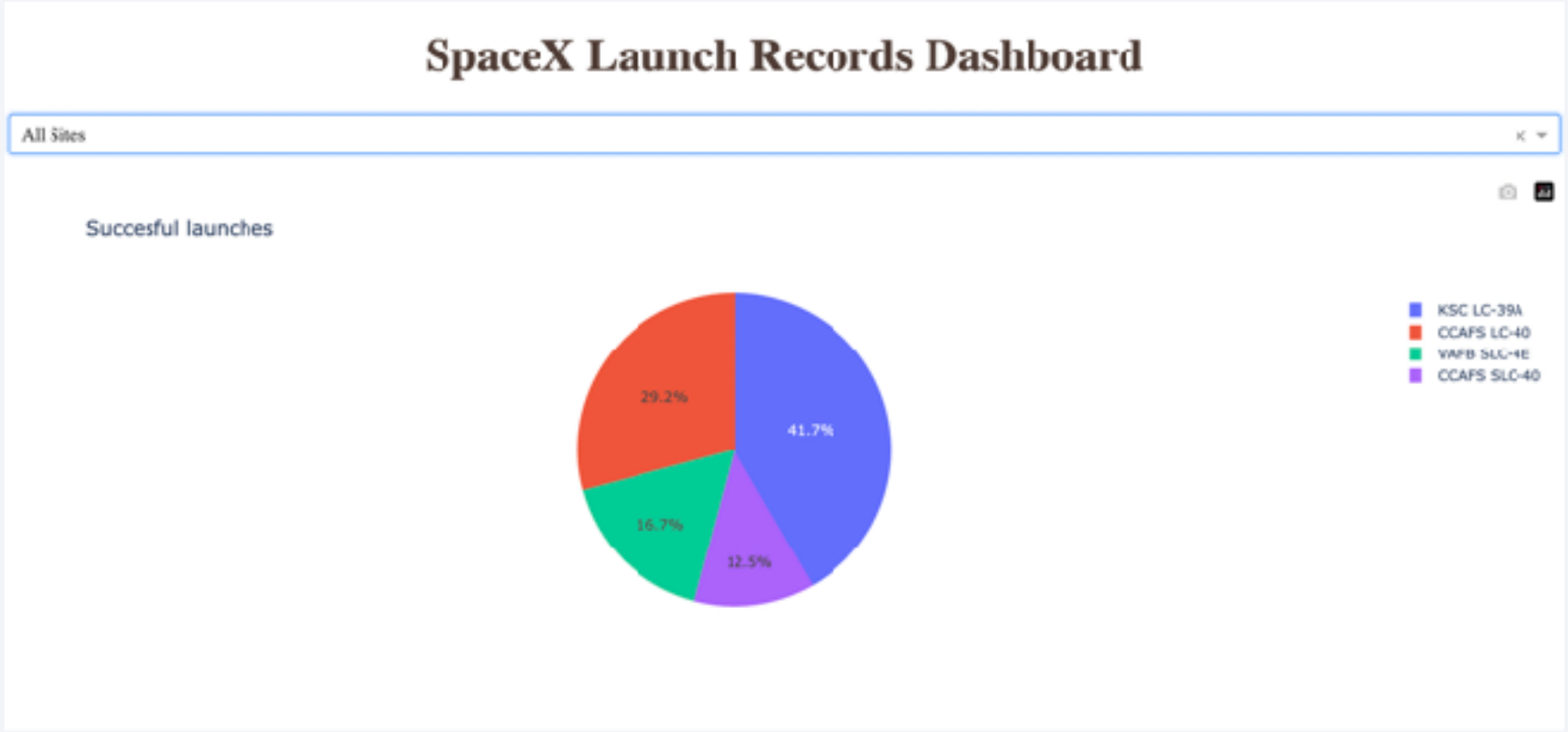


Section 4

Build a Dashboard with Plotly Dash

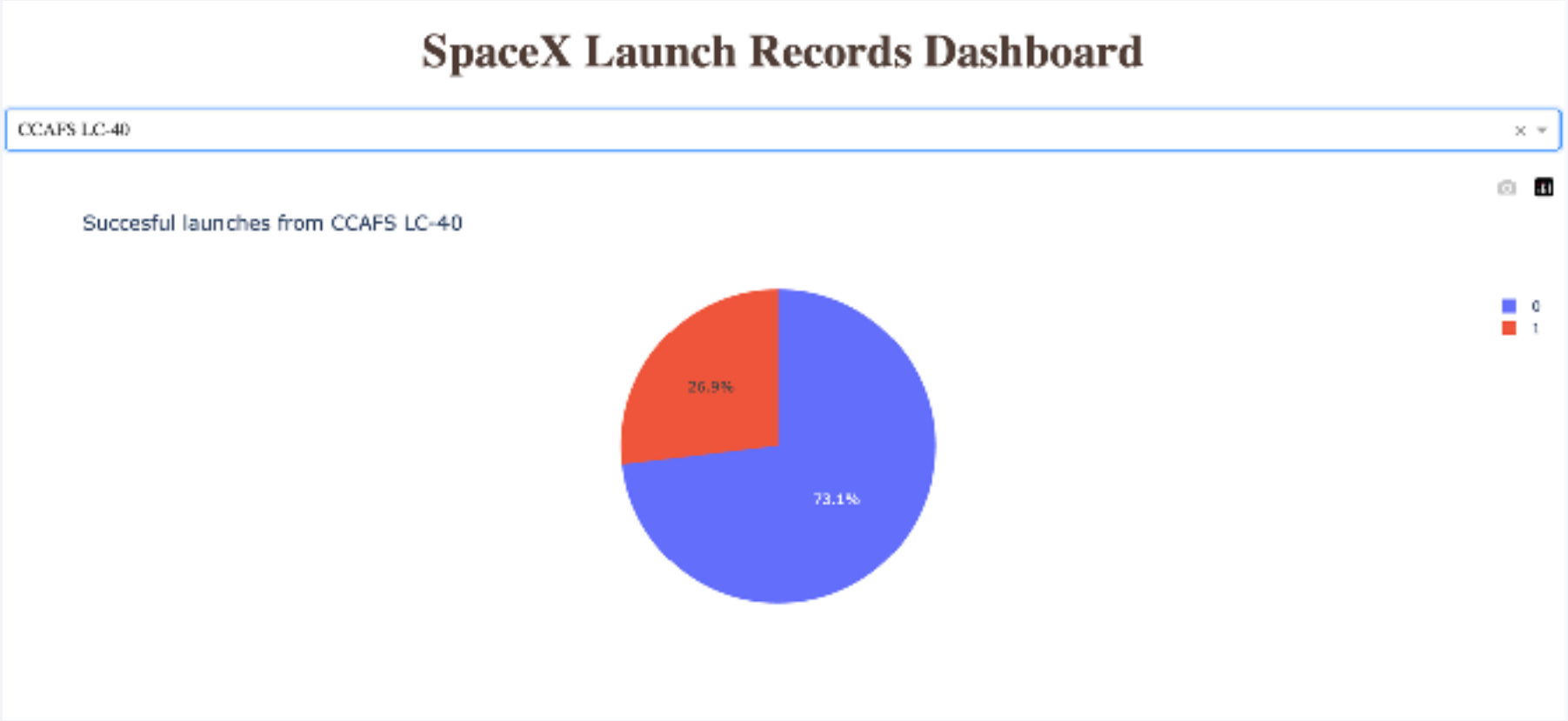
Relative number of successful launches per site

An interactive presentation created with Plotly Dash. Here is presented relative number of successful launches per launch site when all sites are selected.



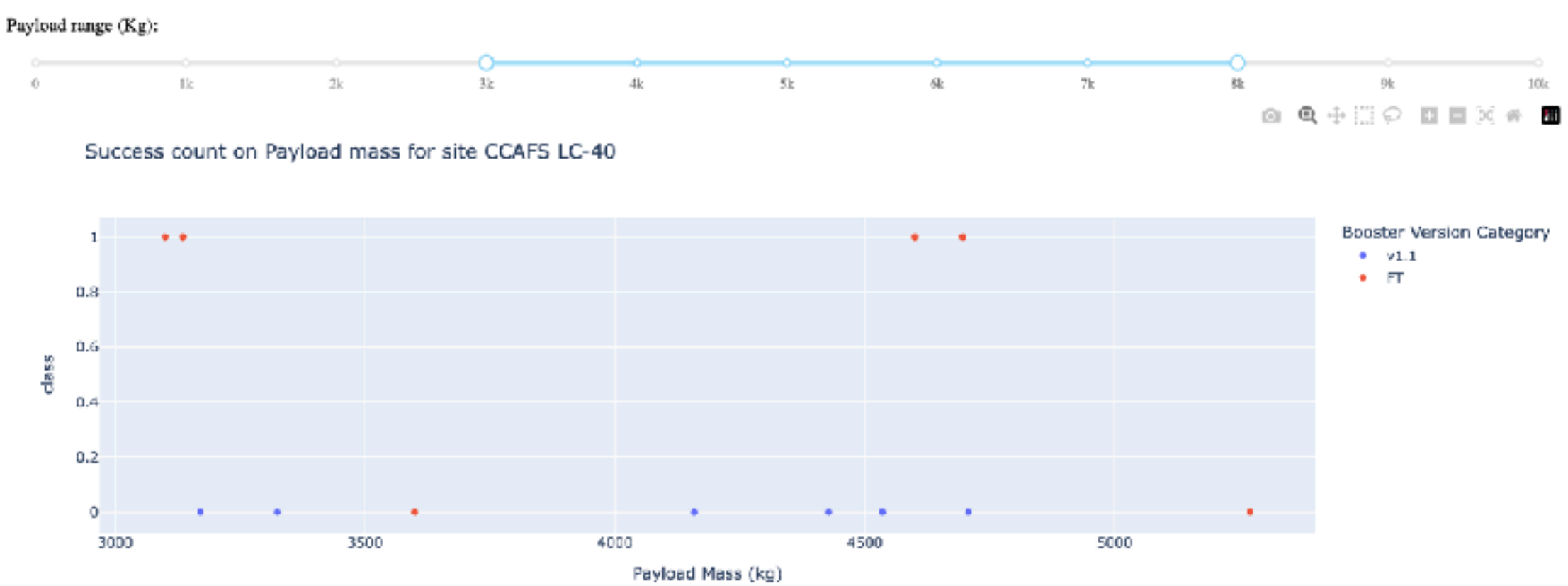
Successful and failed missions for individual launch sites

When a distinct launch site is selected, the chart shows relative numbers of successful and failed missions. Here is presented results for a launch site CCAFS LC-40.



Mission outcomes per payload mass for selected site and a range of payload mass

Here is presented an interactive graph displaying mission outcomes for selected launch sites and payload mass. A range slider was added for selecting a preferred range for the displayed payloads.

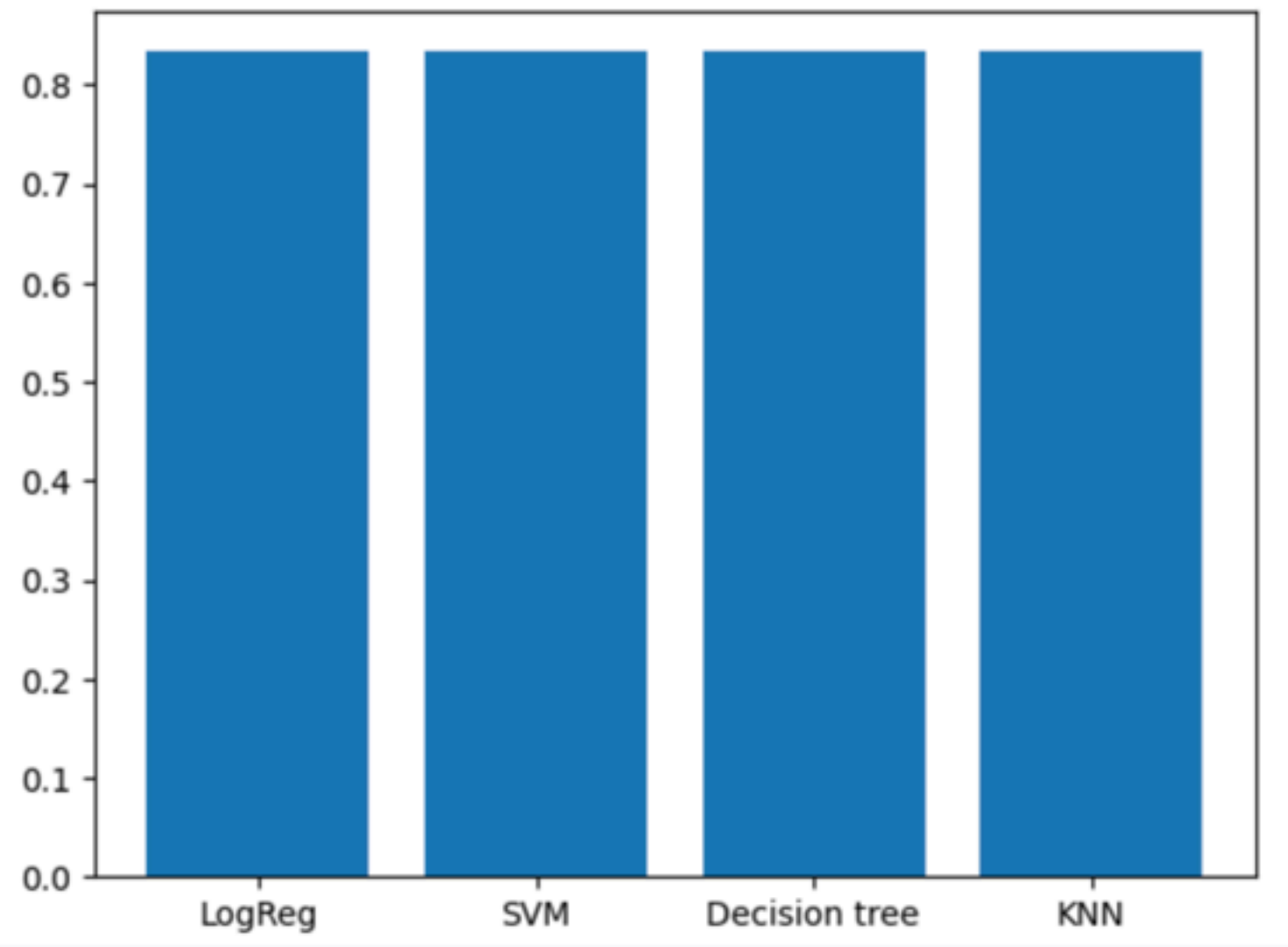


Section 5

Predictive Analysis (Classification)

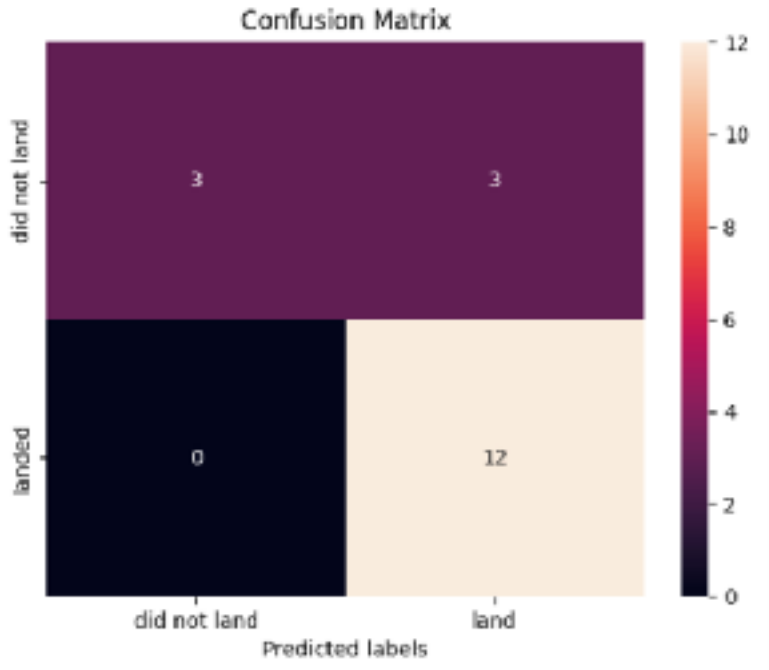
Classification Accuracy

Although models that were tested portrayed different accuracies for the data that was used for training the models, they all had same accuracy of 83,33 % when tested with a set of data that was separate from the data that was used for training.

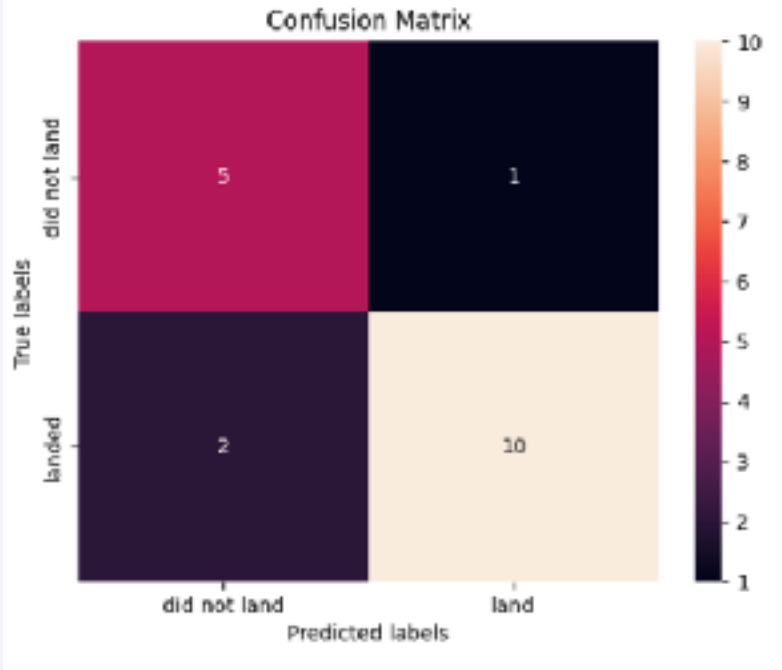


Confusion Matrix

Although the accuracy for the models were same, they gave different predictions as shown in these confusion matrices. The predictions made by decision tree classifier were different when the other three had similar results.



Confusion matrix for logistic regression, support vector machine and k nearest neighbors



Confusion matrix for decision tree classifier

Conclusions

The prediction accuracies were pretty good over all (over 80 %). However, for this relatively small amount of data, no specific selection for the best model can be made.

Confusion matrices do, however, show small differences. When logistic regression, support vector machine and k nearest neighbors didn't have any false negatives, all of them wrongly predicted 3 false positives when using the test data. While decision tree classifier had similarly 3 wrong predictions (hence the same accuracy) they were more evenly distributed between false positives and false negatives (1 and 2, respectively).

This difference might be important when calculating the costs of wrongly predicting a successful landing when the result is actually failure or visa versa.

Thank you!