pplied Data Science capstone

Keshav Hisa

April 18, 202

https://www.google.com/url?sa=i&url=https%3A%2F%2Fspacenews.com%2Fnext-commercial-falcon-heavy-mission-to-launch-debut-astranis-
AOvVaw2Fk66iSj6PnCc2akAiFPco&ust=1641344747041000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCLi42-bzlvUCFQAAAAAdAAAAABAG

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Conclusion

EXECUTIVE SUMMARY

In this capstone project, we will predict if the SpaceX Falcon 9 first stage will land successfully using several machine learning classification algorithms.

The main steps in this project include:

- Data collection, wrangling, and formatting
- Exploratory data analysis
- Interactive data visualization
- Machine learning prediction

Our graphs show that some features of the rocket launches have a correlation with the outcome of the launches, i.e., success or failure.

It is also concluded that decision tree may be the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

INTRODUCTION

In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX 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 SpaceX 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 SpaceX for a rocket launch.

Most unsuccessful landings are planned. Sometimes, SpaceX will perform a controlled landing in the ocean.

The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, launch site, and so on, will the first stage of the rocket land successfully?

METHODOLOGY

The overall methodology includes:

1. Data collection, wrangling, and formatting, using:
   - SpaceX API
   - Web scraping
2. Exploratory data analysis (EDA), using:
   - Pandas and NumPy
   - SQL
3. Data visualization, using:
   - Matplotlib and Seaborn
   - Folium
   - Dash
4. Machine learning prediction, using
   - Logistic regression
   - Support vector machine (SVM)
   - Decision tree
   - K-nearest neighbors (KNN)

METHODOLOGY

1 Data collection, wrangling, and formatting

SpaceX API

- The API used is https://api.spacexdata.com/v4/rockets/.
- The API provides data about many types of rocket launches done by SpaceX, the data is therefore filtered to include only Falcon 9 launches.
- Every missing value in the data is replaced the mean the column that the missing value belongs to.
- We end up with 90 rows or instances and 17 columns or features. The picture below shows the first few rows of the data:

er	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longi
1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577
2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577
3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577
4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610
5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577

METHODOLOGY

1 Data collection, wrangling, and formatting

Web scraping

- The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- The website contains only the data about Falcon 9 launches.
- We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data:

Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	
CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June
CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 Dece
CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May
CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October
CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March

METHODOLOGY

1 Data collection, wrangling, and formatting

The data is later processed so that there are no missing entries and categorical features are encoded using one-hot encoding.

An extra column called ‘Class’ is also added to the data frame. The column ‘Class’ contains 0 if a given launch is failed and 1 if it is successful.

In the end, we end up with 90 rows or instances and 83 columns or features.

METHODOLOGY

2 Exploratory Data Analysis (EDA)

Pandas and NumPy

- Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:
  - The number of launches on each launch site
  - The number of occurrence of each orbit
  - The number and occurrence of each mission outcome

SQL

- The data is queried using SQL to answer several questions about the data such as:
  - 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

METHODOLOGY

3 Data Visualization

Matplotlib and Seaborn

- Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts.
- The plots and charts are used to understand more about the relationships between several features, such as:
  - The relationship between flight number and launch site
  - The relationship between payload mass and launch site
  - The relationship between success rate and orbit type

Folium

- Functions from the Folium libraries are used to visualize the data through interactive maps.
- The Folium library is used to:
  - Mark all launch sites on a map
  - Mark the succeeded launches and failed launches for each site on the map
  - Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

METHODOLOGY

3 Data Visualization

Dash

- Functions from Dash are used to generate an interactive site where we can toggle the input using a dropdown menu and a range slider.
- Using a pie chart and a scatterplot, the interactive site shows:
  - The total success launches from each launch site
  - The correlation between payload mass and mission outcome (success or failure) for each launch site

METHODOLOGY

4 Machine Learning Prediction

Functions from the Scikit-learn library are used to create our machine learning models.

The machine learning prediction phase include the following steps:

- Standardizing the data
- Splitting the data into training and test data
- Creating machine learning models, which include:
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K nearest neighbors (KNN)
- Fit the models on the training set
- Find the best combination of hyperparameters for each model
- Evaluate the models based on their accuracy scores and confusion matrix

RESULTS

The results are split into 5 sections:

- SQL (EDA with SQL)
- Matplotlib and Seaborn (EDA with Visualization)
- Folium
- Dash
- Predictive Analysis

In all of the graphs that follow, class 0 represents a failed launch outcome while class 1 represents a successful launch outcome.

RESULTS

1 SQL (EDA with SQL)

The names of the unique launch sites in the space mission

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

5 records where launch sites begin with 'CCA'

ne__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	l
18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	F
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	F
07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	
00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	
15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	

RESULTS

1 SQL (EDA with SQL)

The total payload mass carried by boosters launched by NASA (CRS)

Total payload mass by NASA (CRS)
45596

The average payload mass carried by booster version F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

The date when the first successful landing outcome in ground pad was achieved

Date of first successful landing outcome in ground pad
2015-12-22

RESULTS

1 SQL (EDA with SQL)

The names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

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

The total number of successful and failure mission outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1

RESULTS

1 SQL (EDA with SQL)

The names of the booster versions which have carried the maximum payload mass

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

RESULTS
1 SQL (EDA with SQL)

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

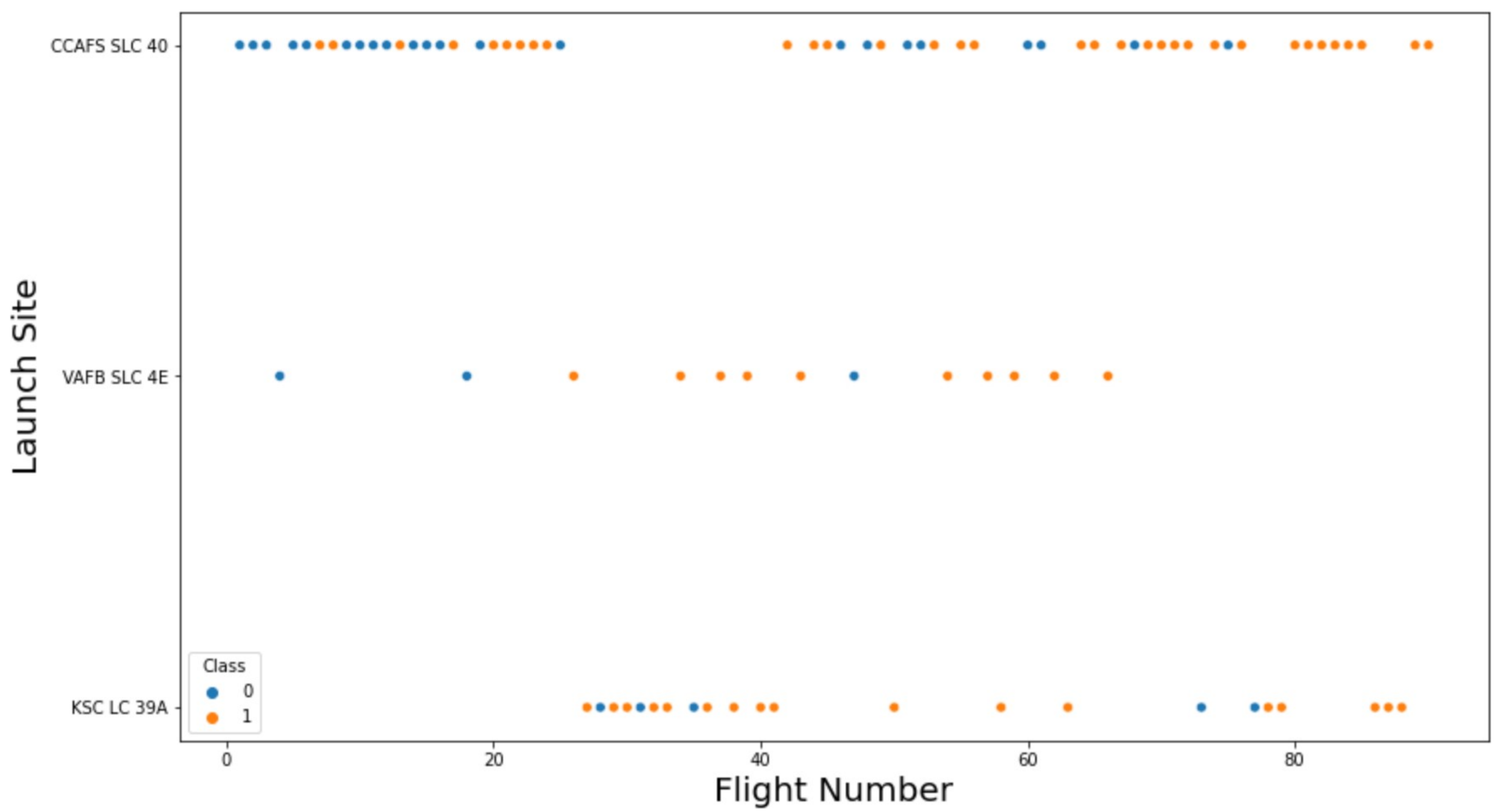
The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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

RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

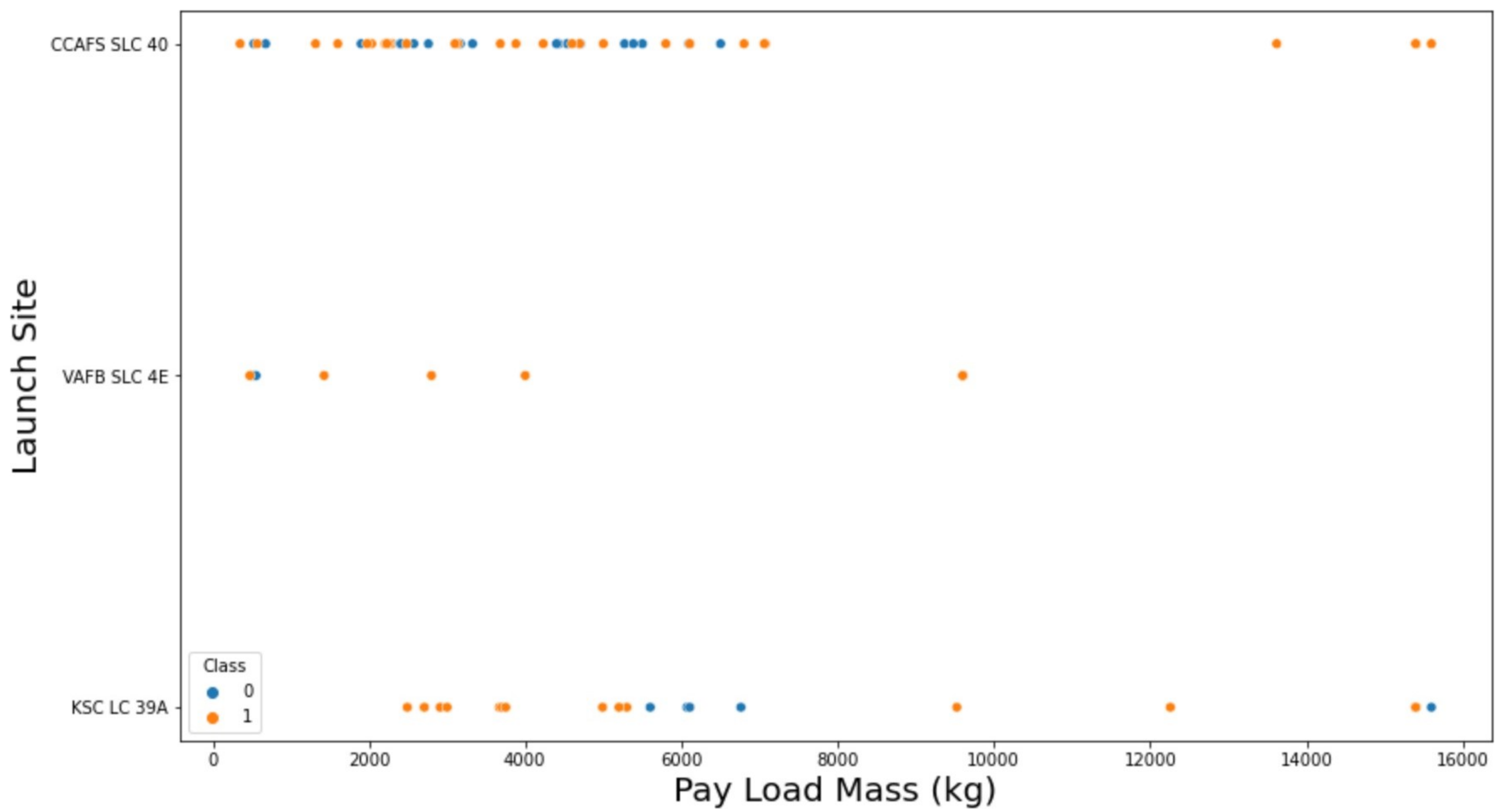
The relationship between flight number and launch site



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

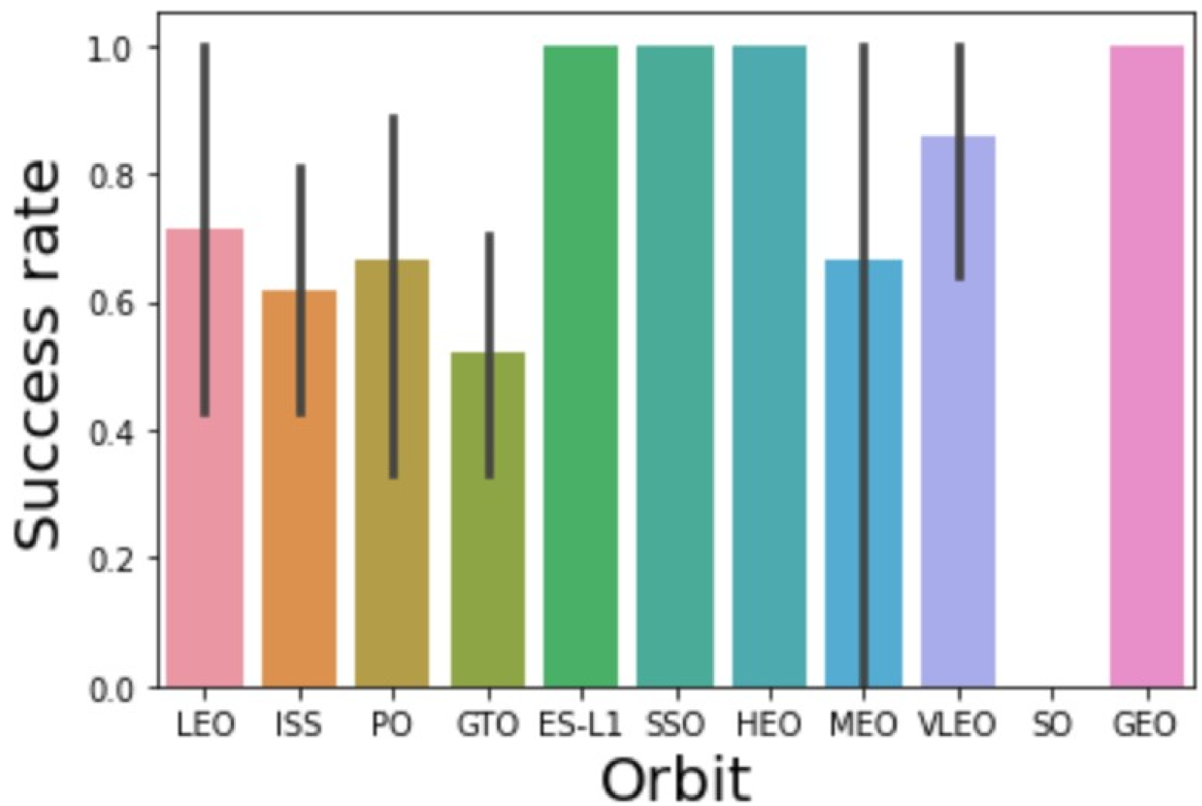
The relationship between payload mass and launch site



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

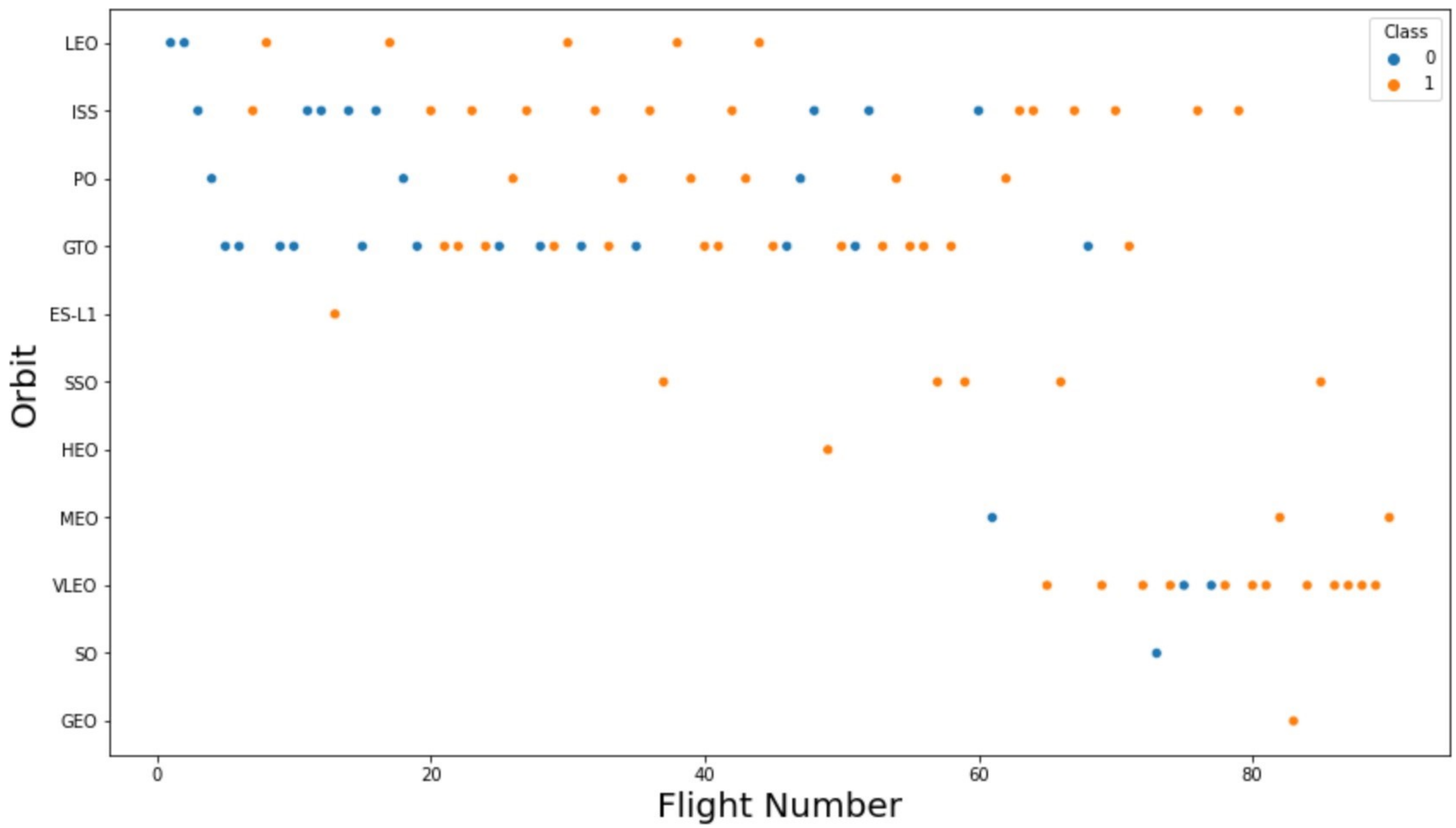
The relationship between success rate and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

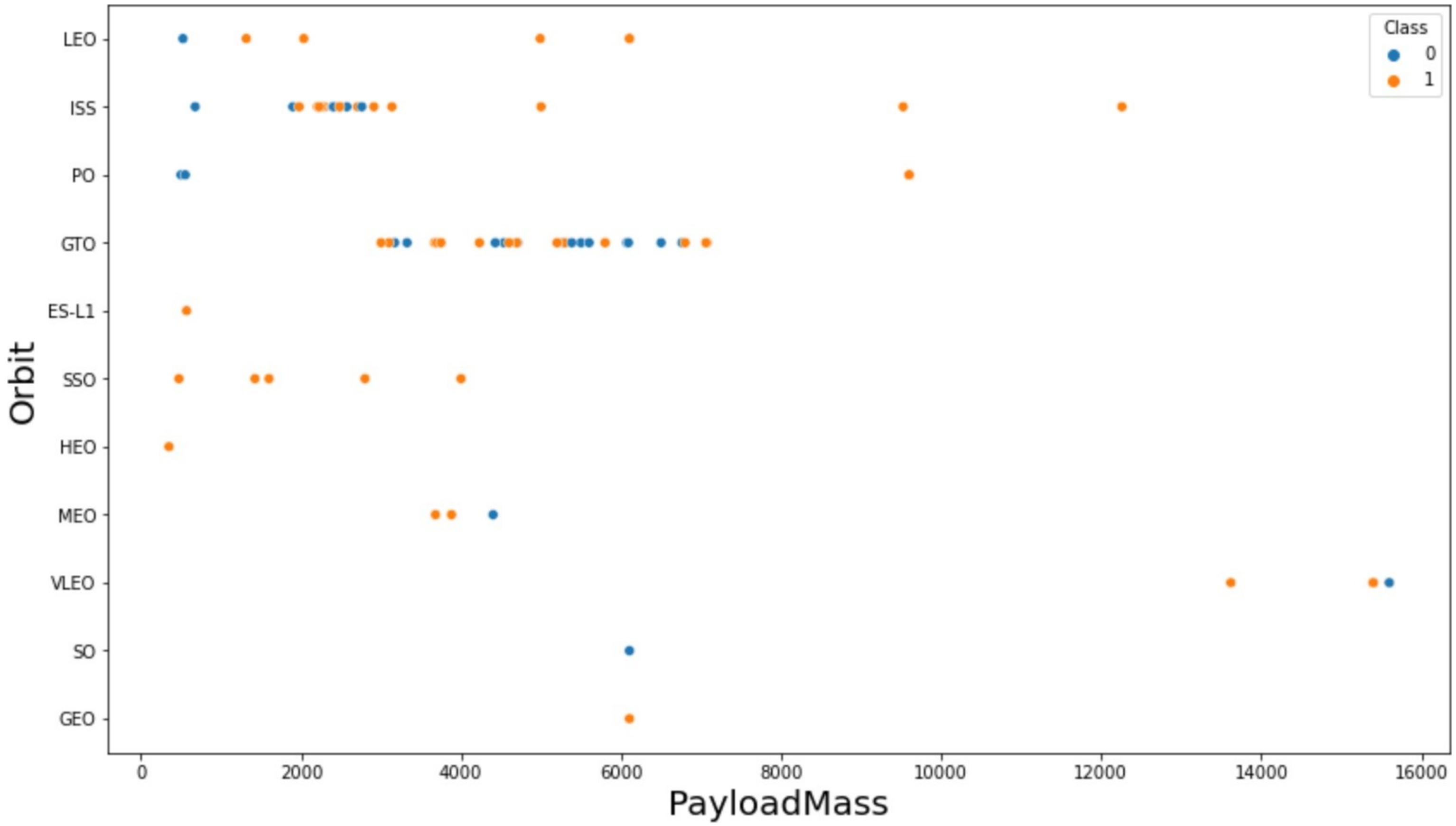
The relationship between flight number and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

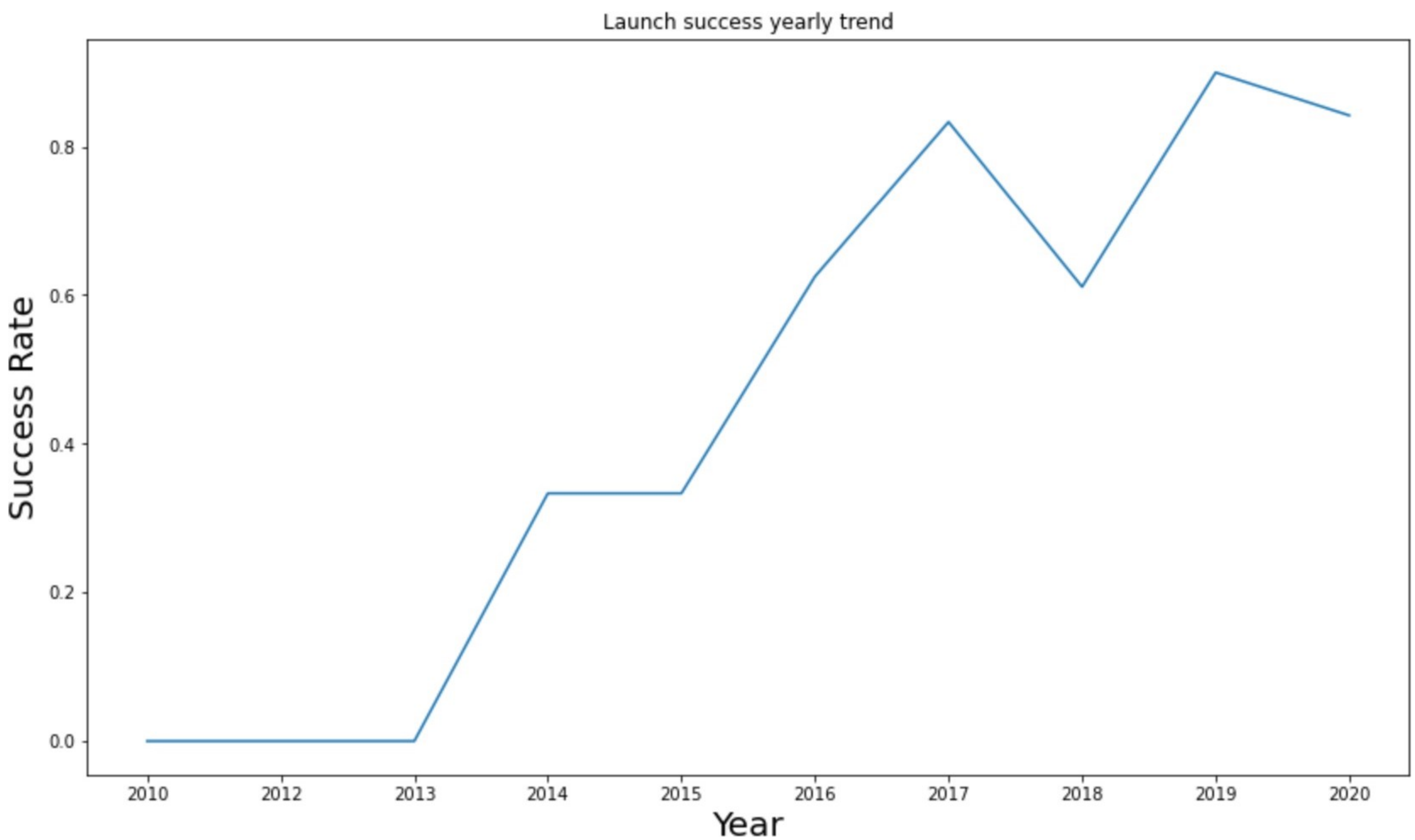
The relationship between payload mass and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

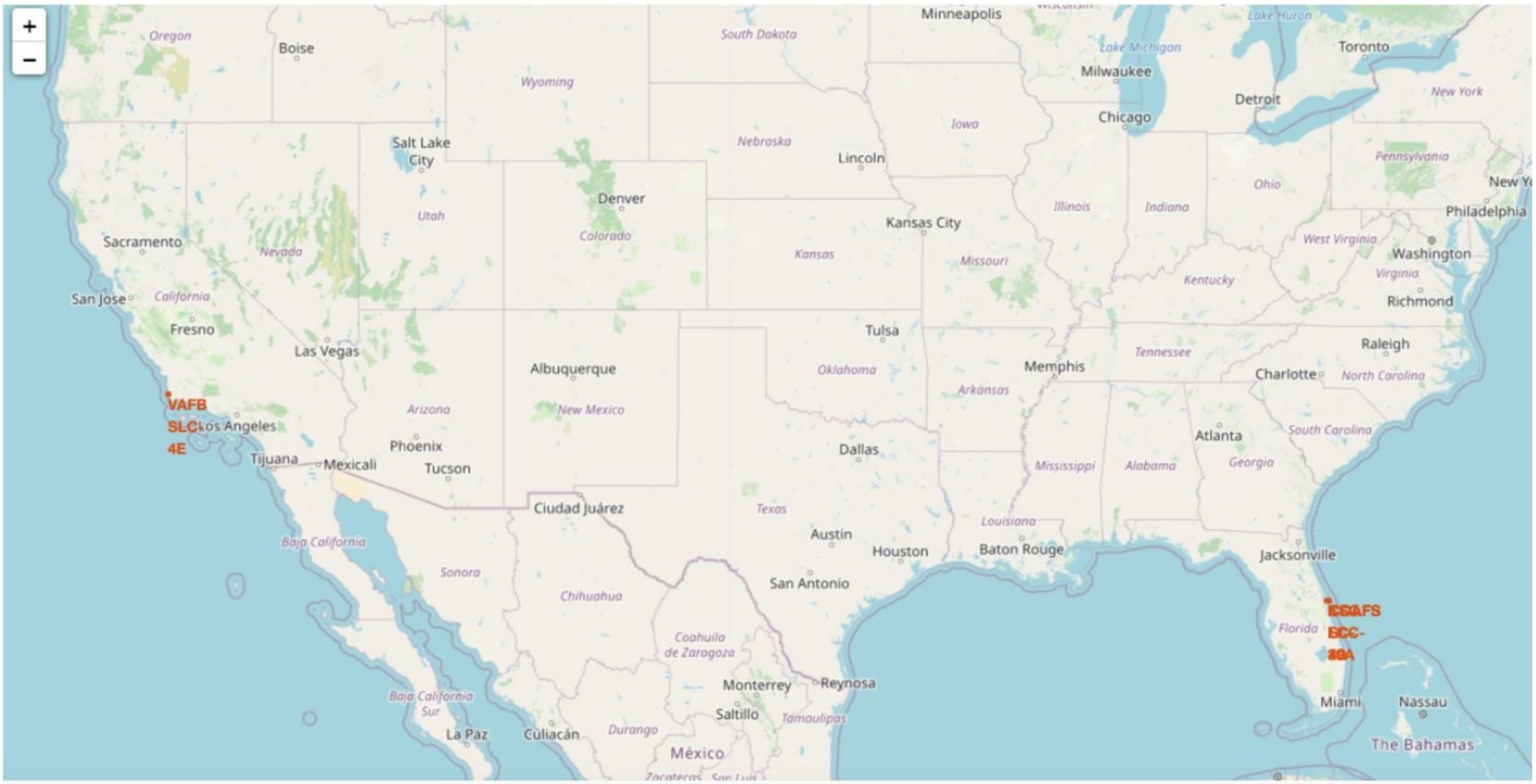
The launch success yearly trend



RESULTS

3 Folium

All launch sites on map

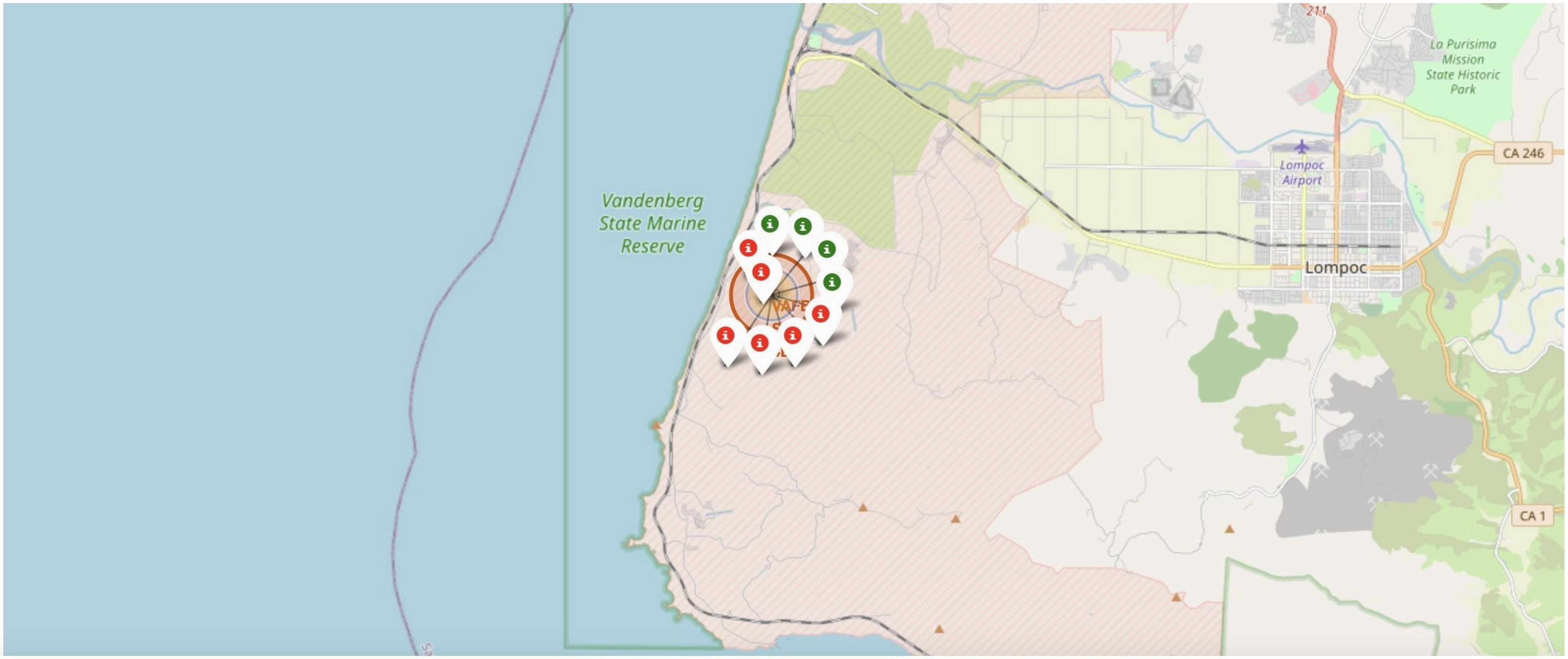


RESULTS

3 Folium

The succeeded launches and failed launches for each site on map

- If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch

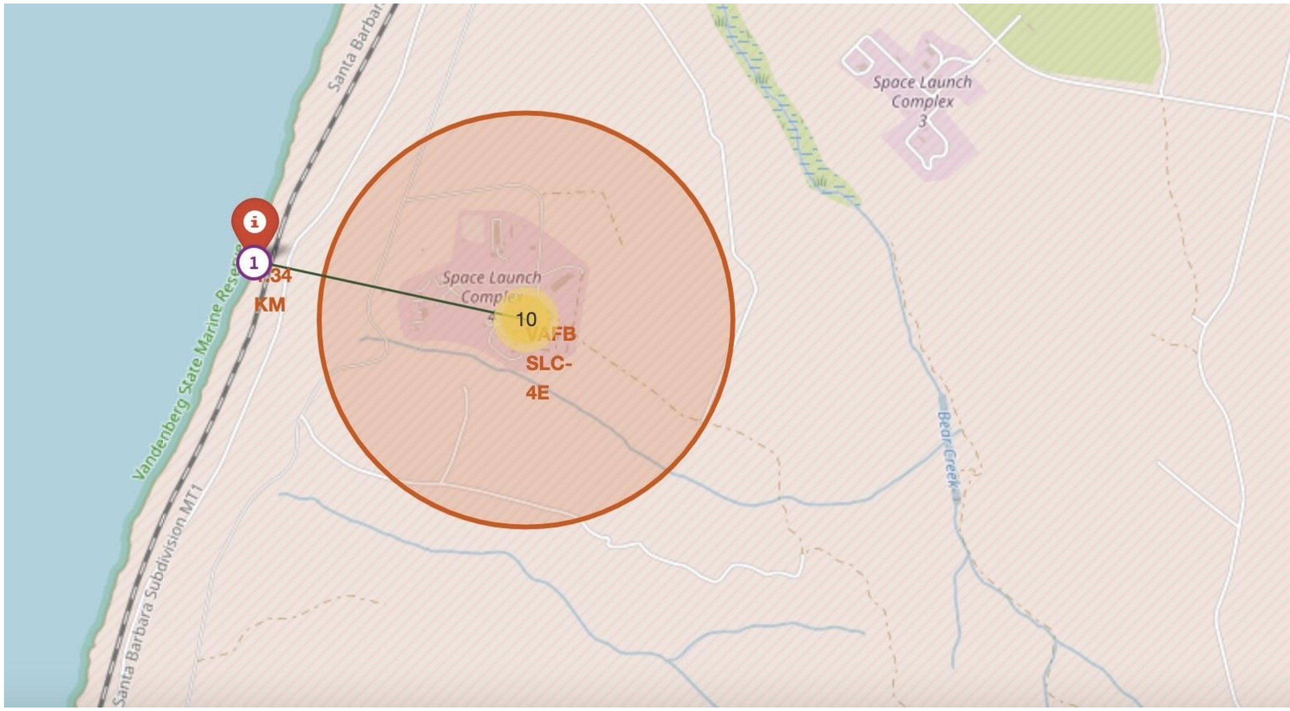


RESULTS

3 Folium

The distances between a launch site to its proximities such as the nearest city, railway, or highway

- The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline

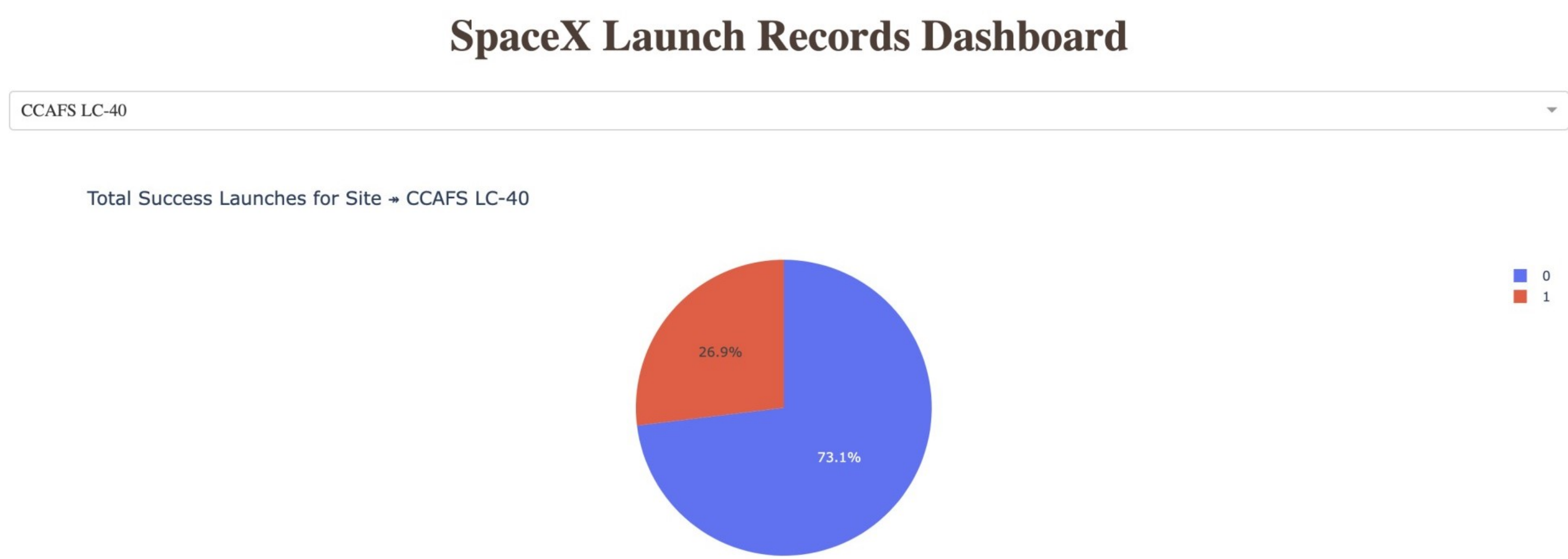


RESULTS

4 Dash

The picture below shows a pie chart when launch site CCAFS LC-40 is chosen.

0 represents failed launches while 1 represents successful launches. We can see that 73.1% of launches done at CCAFS LC-40 are failed launches.

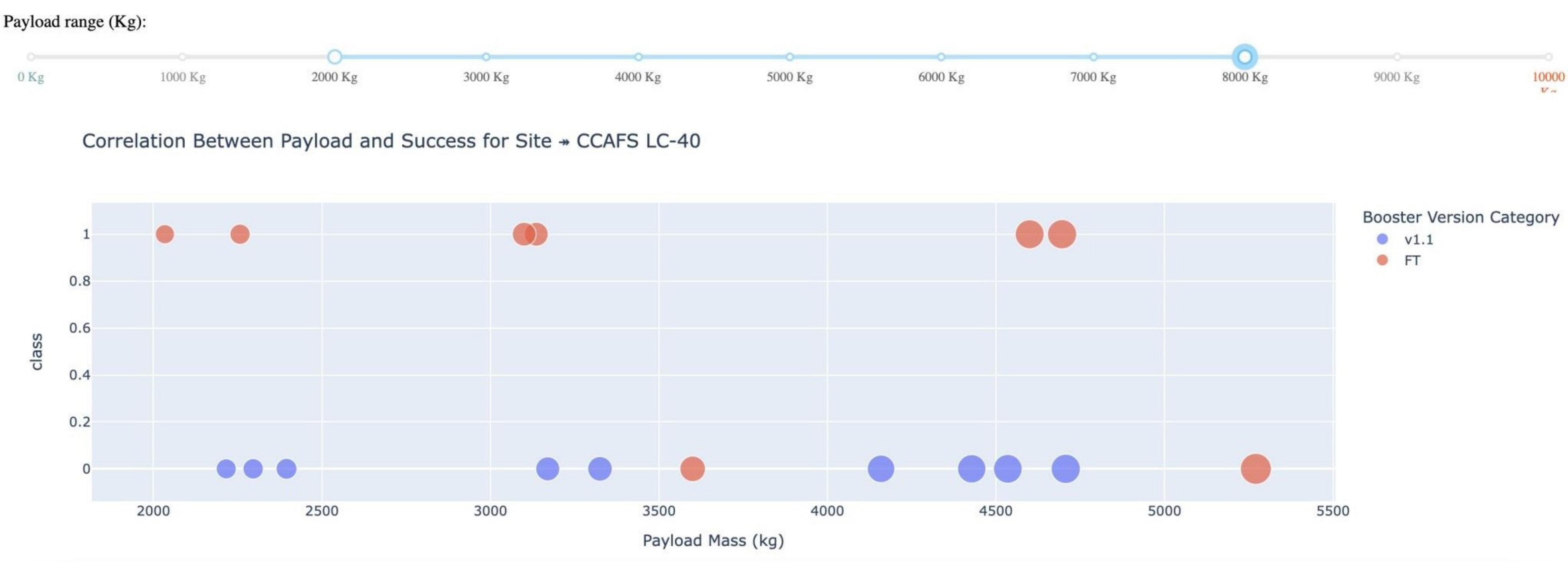


RESULTS

4 Dash

The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.

Class 0 represents failed launches while class 1 represents successful launches.

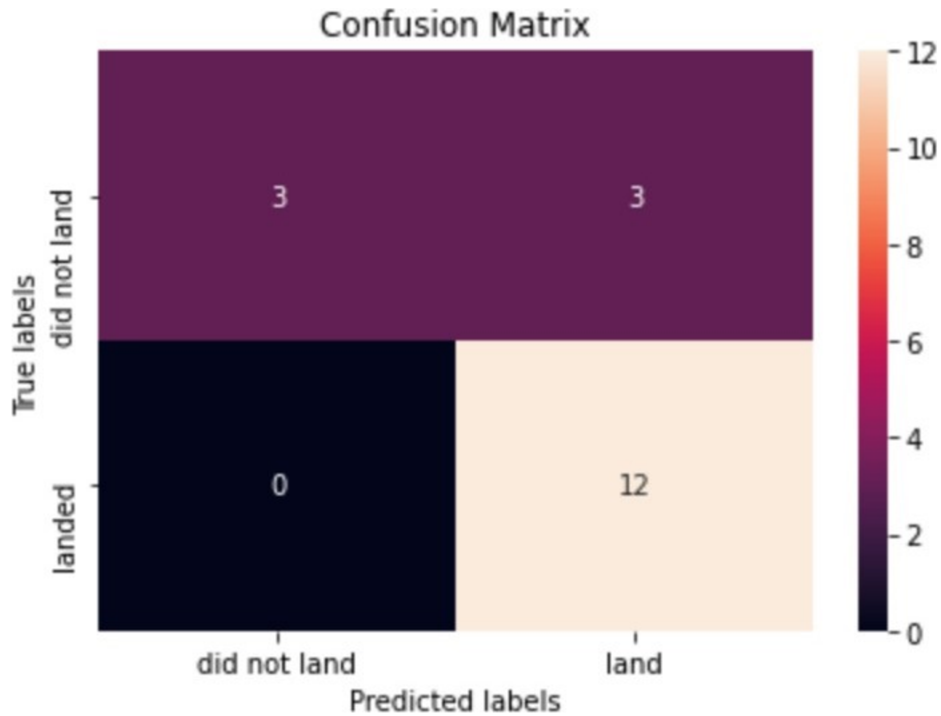


RESULTS

5 Predictive Analysis

Logistic regression

- GridSearchCV best score: 0.8464285714285713
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:

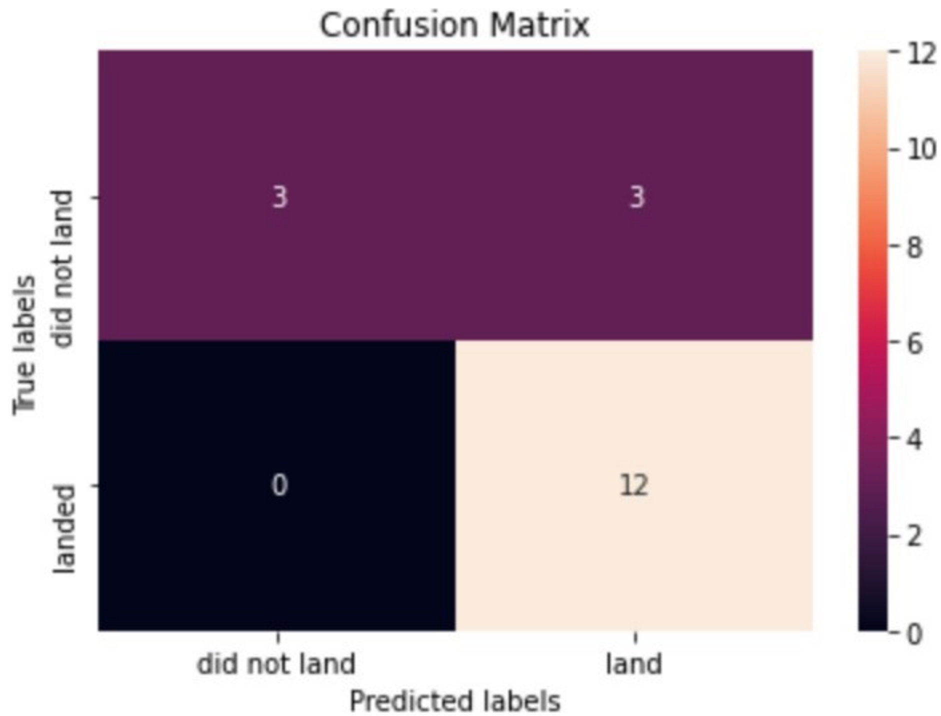


RESULTS

5 Predictive Analysis

Support vector machine (SVM)

- GridSearchCV best score: 0.8482142857142856
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:

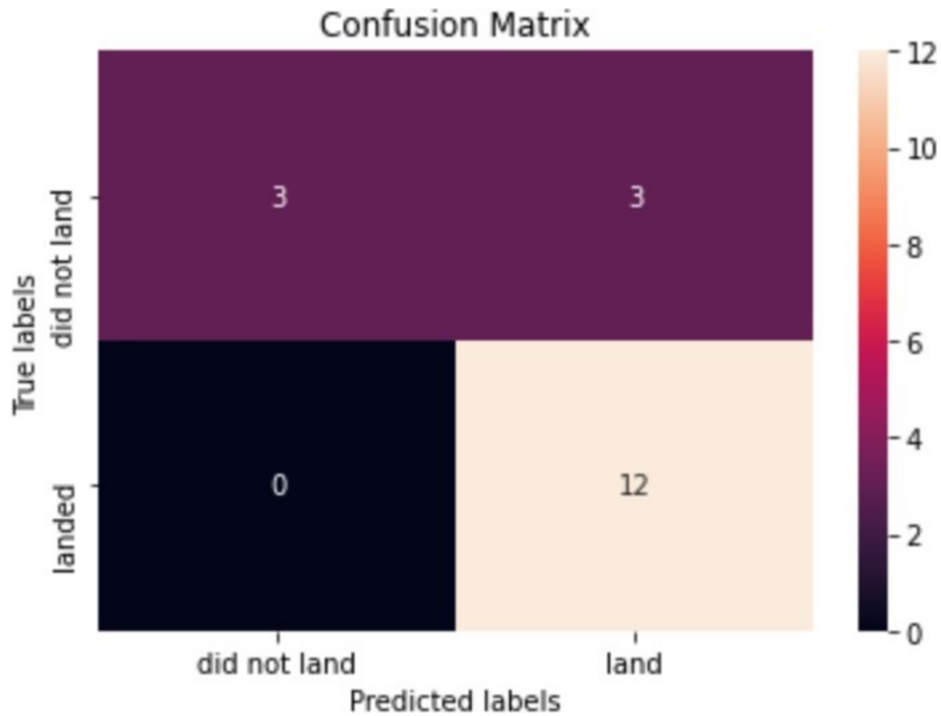


RESULTS

5 Predictive Analysis

Decision tree

- GridSearchCV best score: 0.8892857142857142
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:

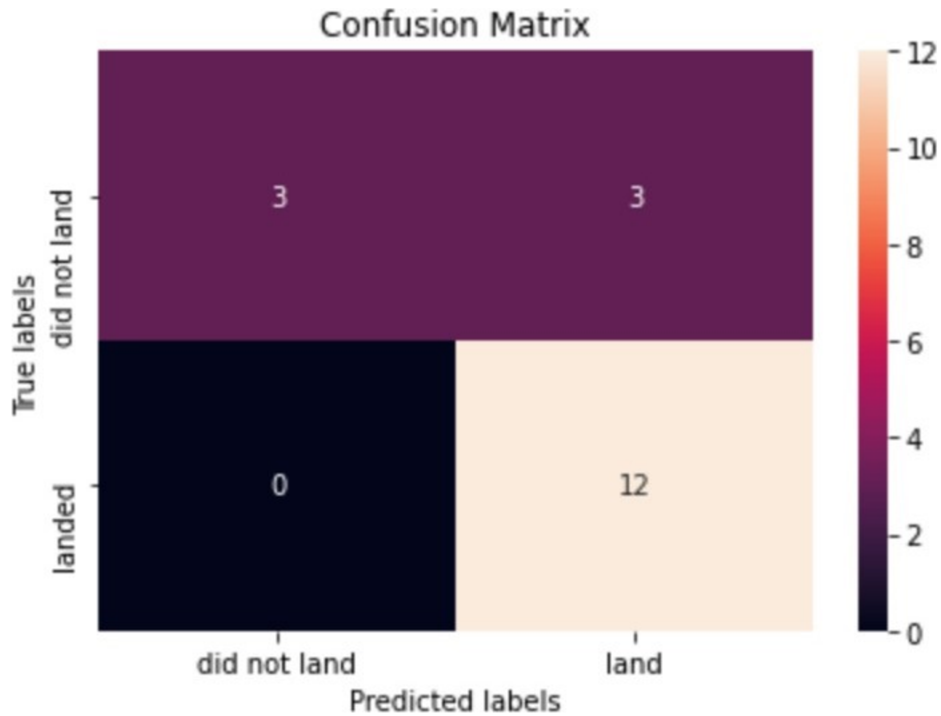


RESULTS

5 Predictive Analysis

K nearest neighbors (KNN)

- GridSearchCV best score: 0.8482142857142858
- Accuracy score on test set: 0.8333333333333334
- Confusion matrix:



RESULTS

5 Predictive Analysis

Putting the results of all 4 models side by side, we can see that they all share the same accuracy score and confusion matrix when tested on the test set.

Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:

1. Decision tree (GridSearchCV best score: 0.8892857142857142)
2. K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
3. Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
4. Logistic regression (GridSearchCV best score: 0.8464285714285713)

DISCUSSION

From the data visualization section, we can see that some features may have correlation with the mission outcome in several ways. For example, with heavy payloads the successful landing or positive landing rate are more for orbit types Polar, LEO and ISS. However, for GTO, we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.

Therefore, each feature may have a certain impact on the final mission outcome. The exact ways of how each of these features impact the mission outcome are difficult to decipher. However, we can use some machine learning algorithms to learn the pattern of the past data and predict whether a mission will be successful or not based on the given features.

CONCLUSION

In this project, we try to predict if the first stage of a given Falcon 9 launch will land in order to determine the cost of a launch.

Each feature of a Falcon 9 launch, such as its payload mass or orbit type, may affect the mission outcome in a certain way.

Several machine learning algorithms are employed to learn the patterns of past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.

The predictive model produced by decision tree algorithm performed the best among the 4 machine learning algorithms employed.