✖

Applied Data Science capstone

AniKet Gusain

CONTENT

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

SUMMARY

- In this capstone project, we will use machine learning classification methods to predict whether the SpaceX Falcon 9 first stage will successfully land.
- The main steps in this project include:
  - Data collection, wrangling, and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- Our graphs illustrate that several aspects of rocket launches are related to the success or failure of the launches.
- It is also found that the decision tree is the best machine learning technique for predicting whether the Falcon 9 first stage will successfully land.

INTRODUCTION

- In this capstone, we will predict if the Falcon 9 first stage will successfully land. SpaceX advertises Falcon 9 rocket launches on its website for 62 million dollars; other companies charge up to 165 million dollars apiece; much of the savings is due to SpaceX's ability to reuse the first stage. As a result, if we can predict whether the first stage will land, we can estimate the cost of a launch. This data can be used if another company wishes to compete with SpaceX for a rocket launch.
- The majority of failed landings are planned. SpaceX will occasionally perform a controlled landing in the water.
- The fundamental issue we are attempting to answer is whether, for a given set of Falcon 9 rocket launch features such as payload mass, orbit type, launch site, and so on, the first stage of the rocket will successfully land.

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:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

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:

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

METHODOLOGY

1 Data collection, wrangling, and formatting

- The data is then examined to ensure that no entries are missing, 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
- Class 0 denotes a failed launch outcome, whereas class 1 represents a successful launch outcome in all of the graphs that follow.

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'

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	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

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

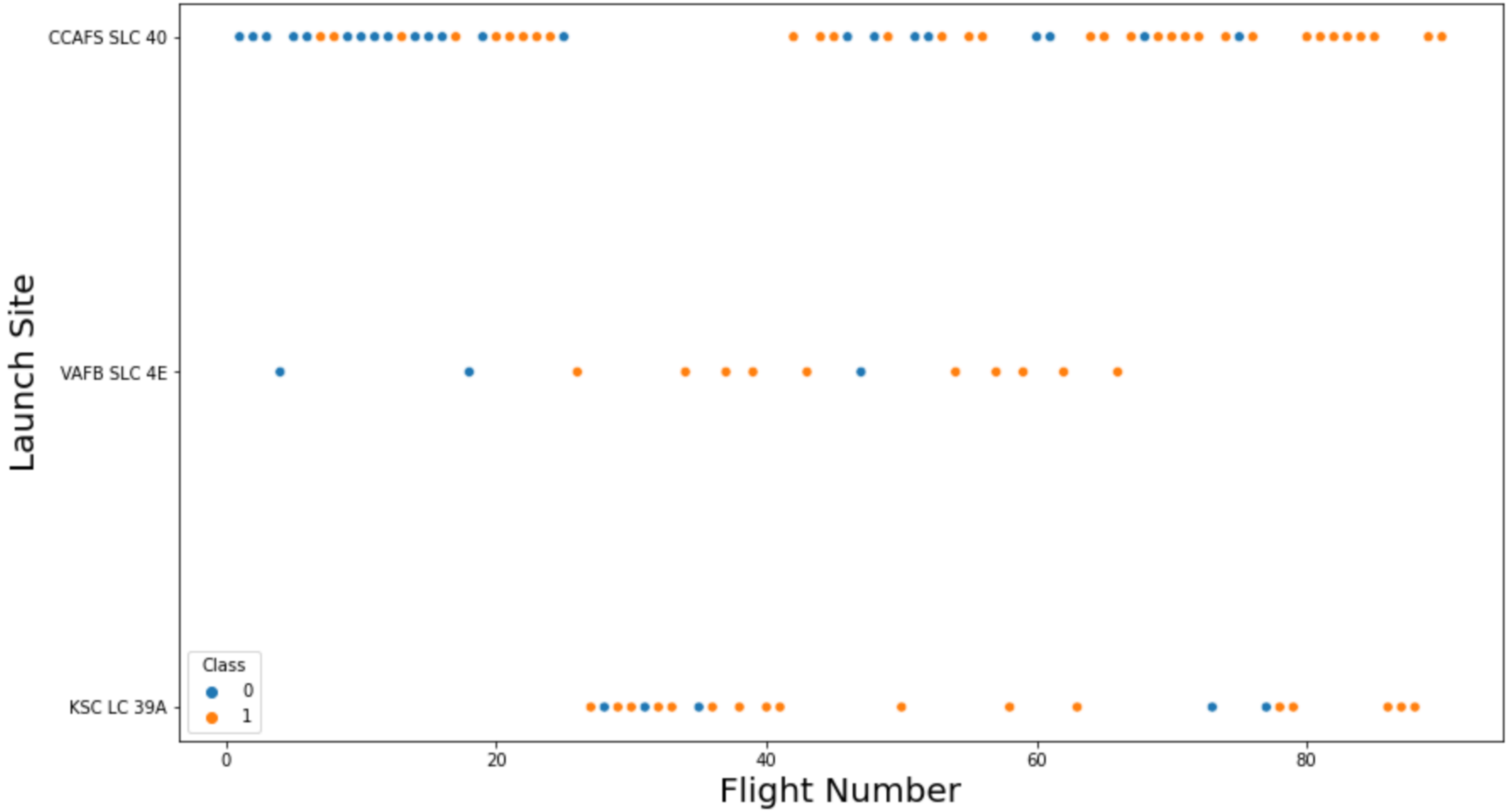
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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)

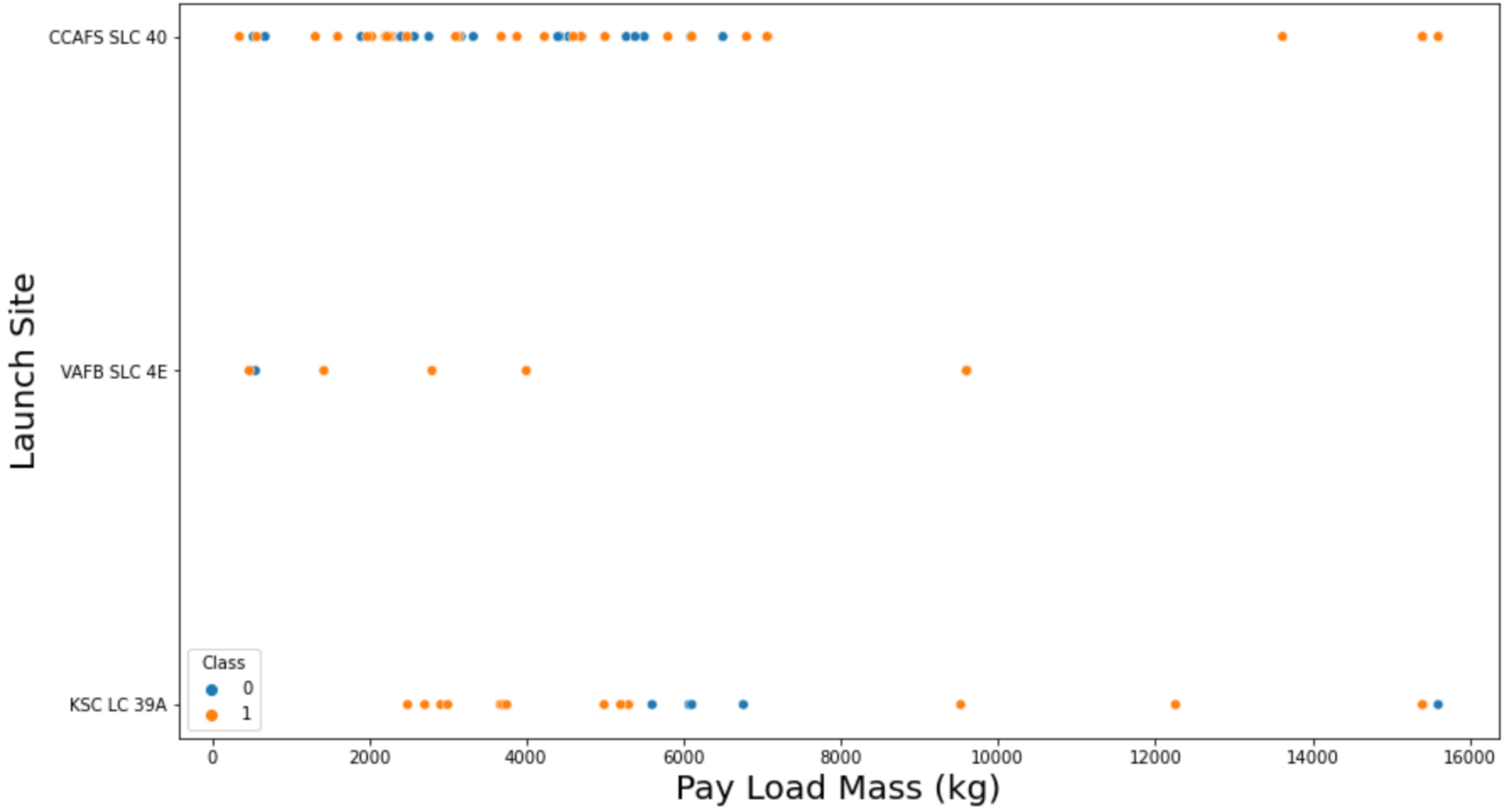
- The relationship between flight number and launch site



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

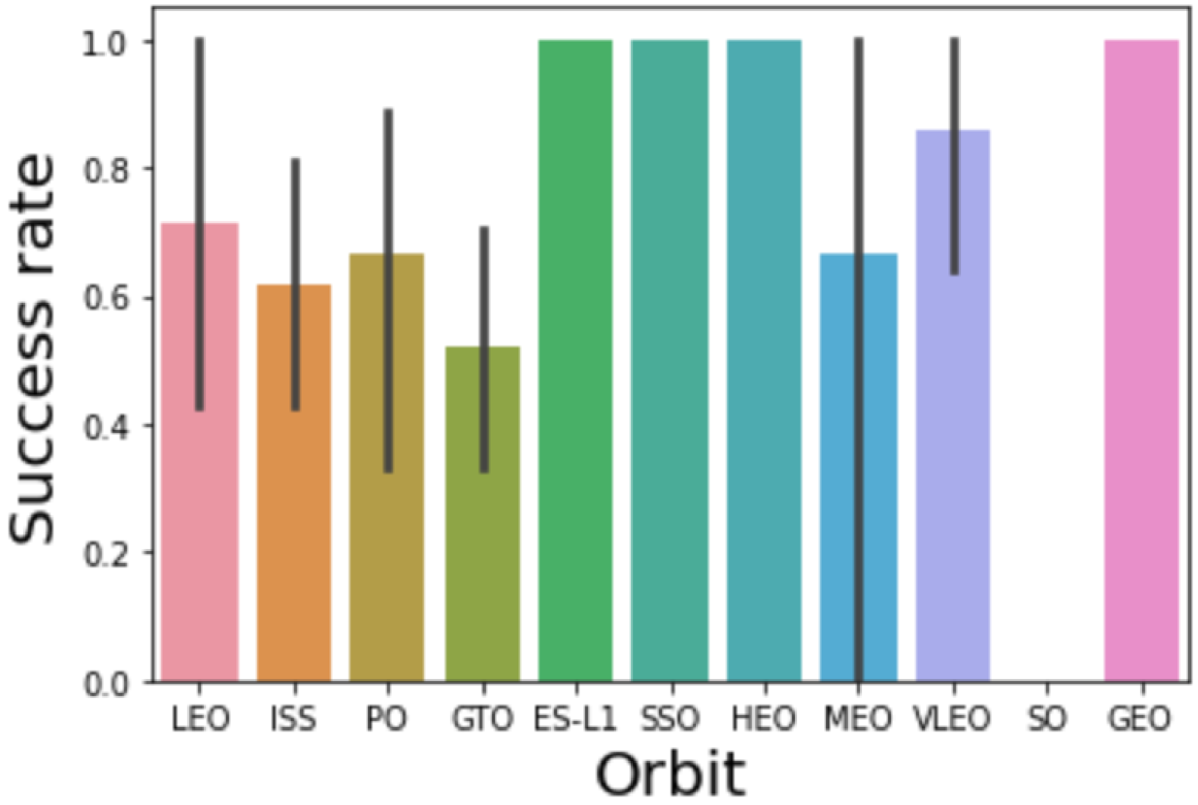
- The relationship between payload mass and launch site



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

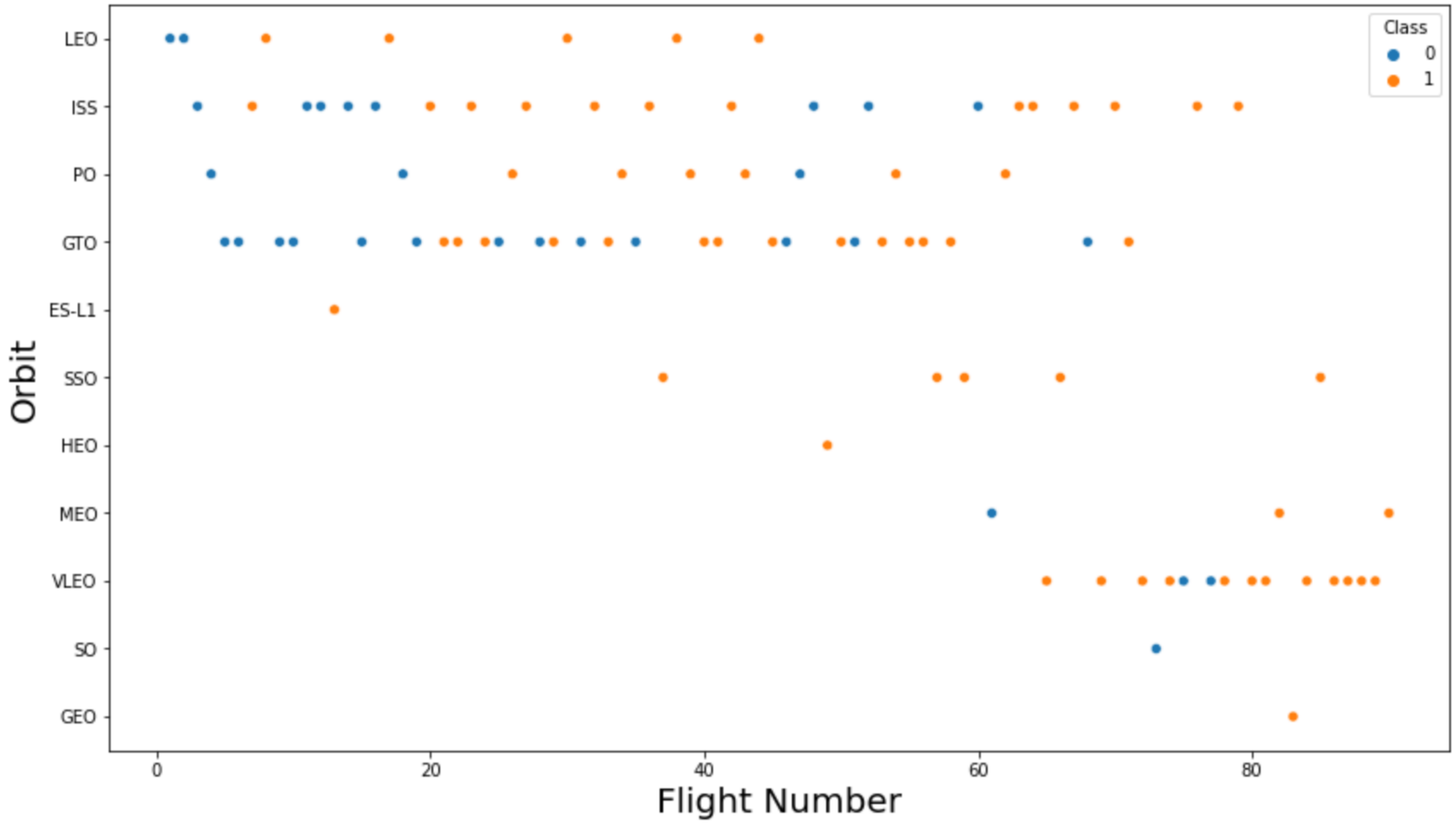
- The relationship between success rate and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

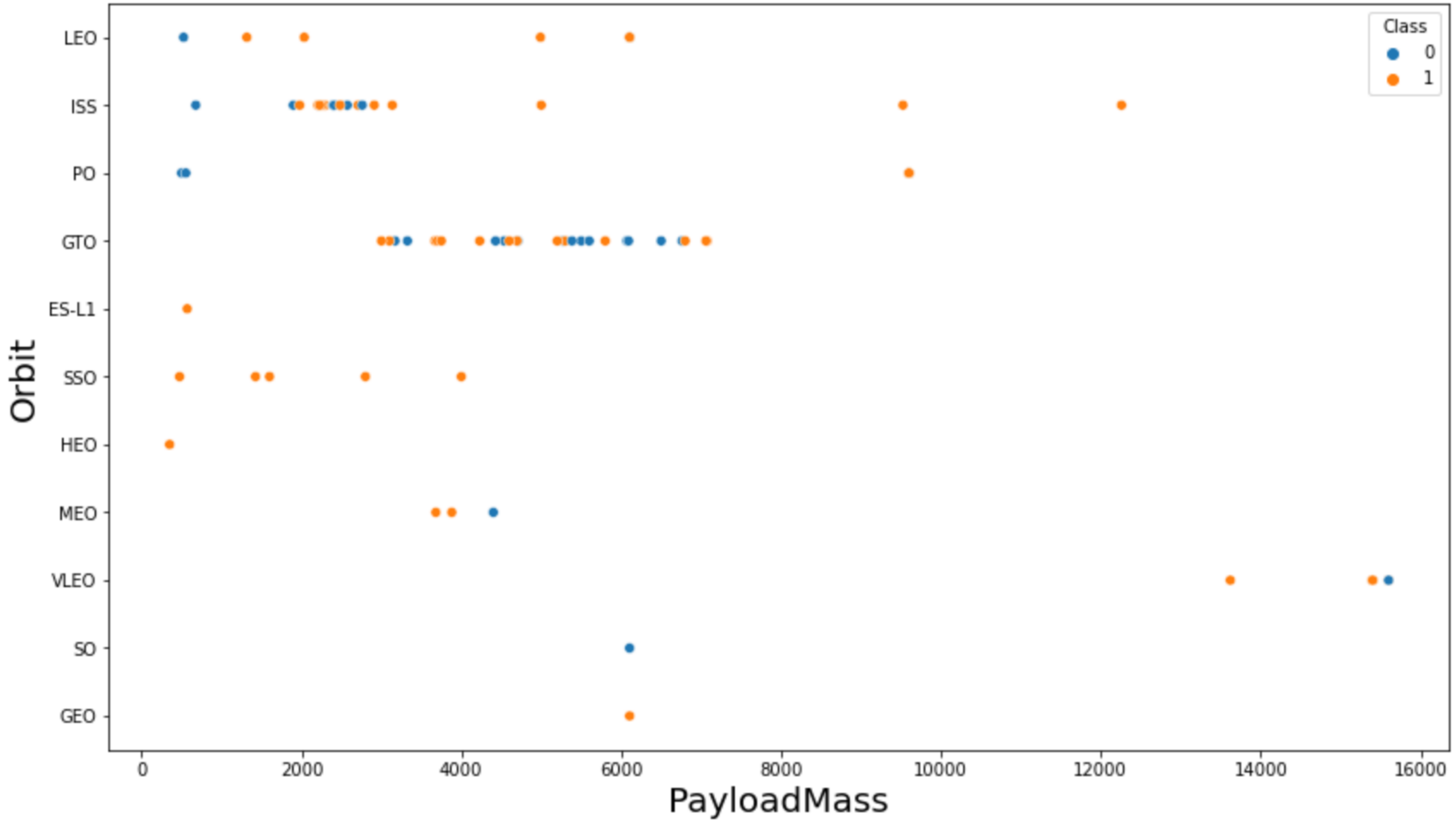
- The relationship between flight number and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

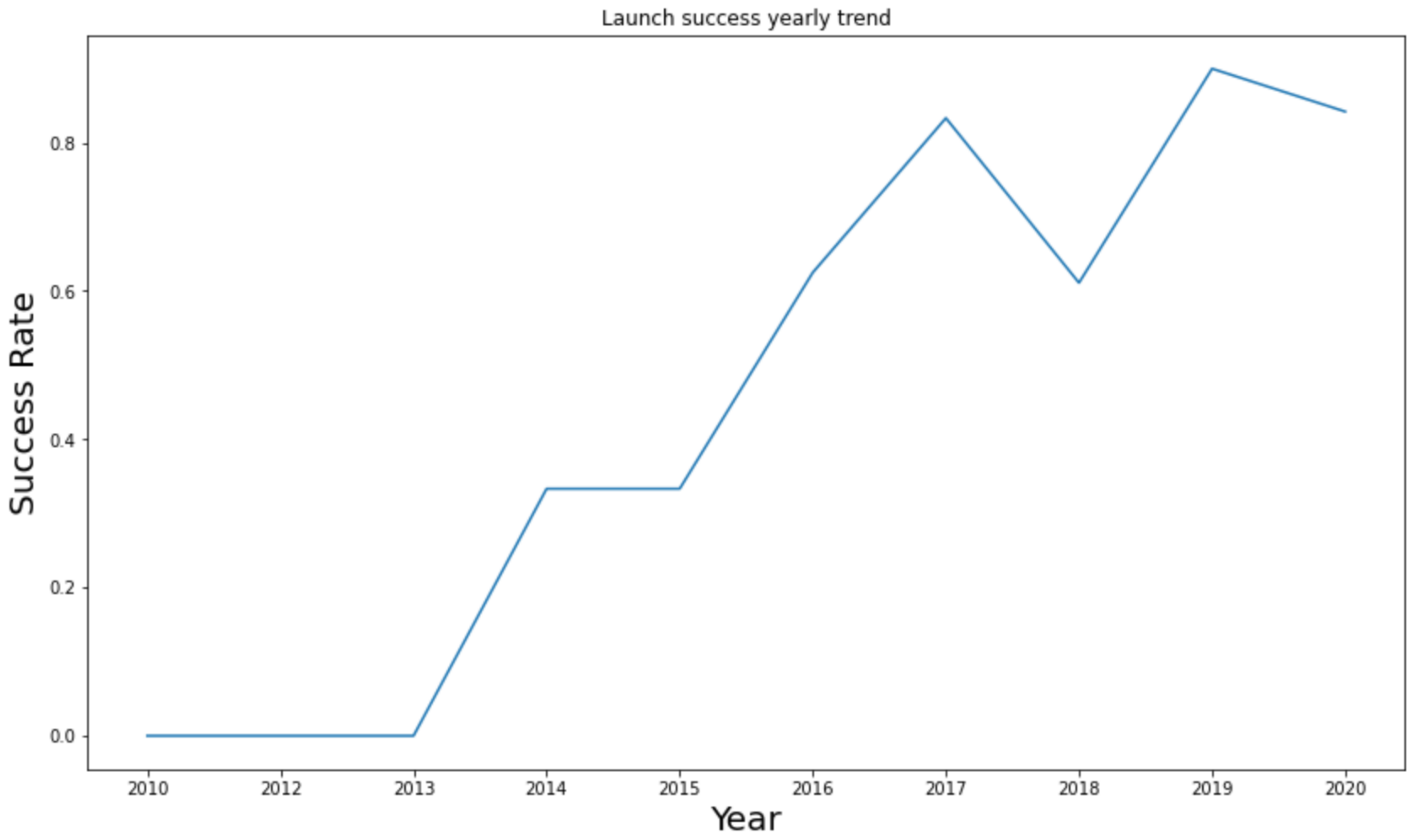
- The relationship between payload mass and orbit type



RESULTS

2 Matplotlib and Seaborn (EDA with Visualization)

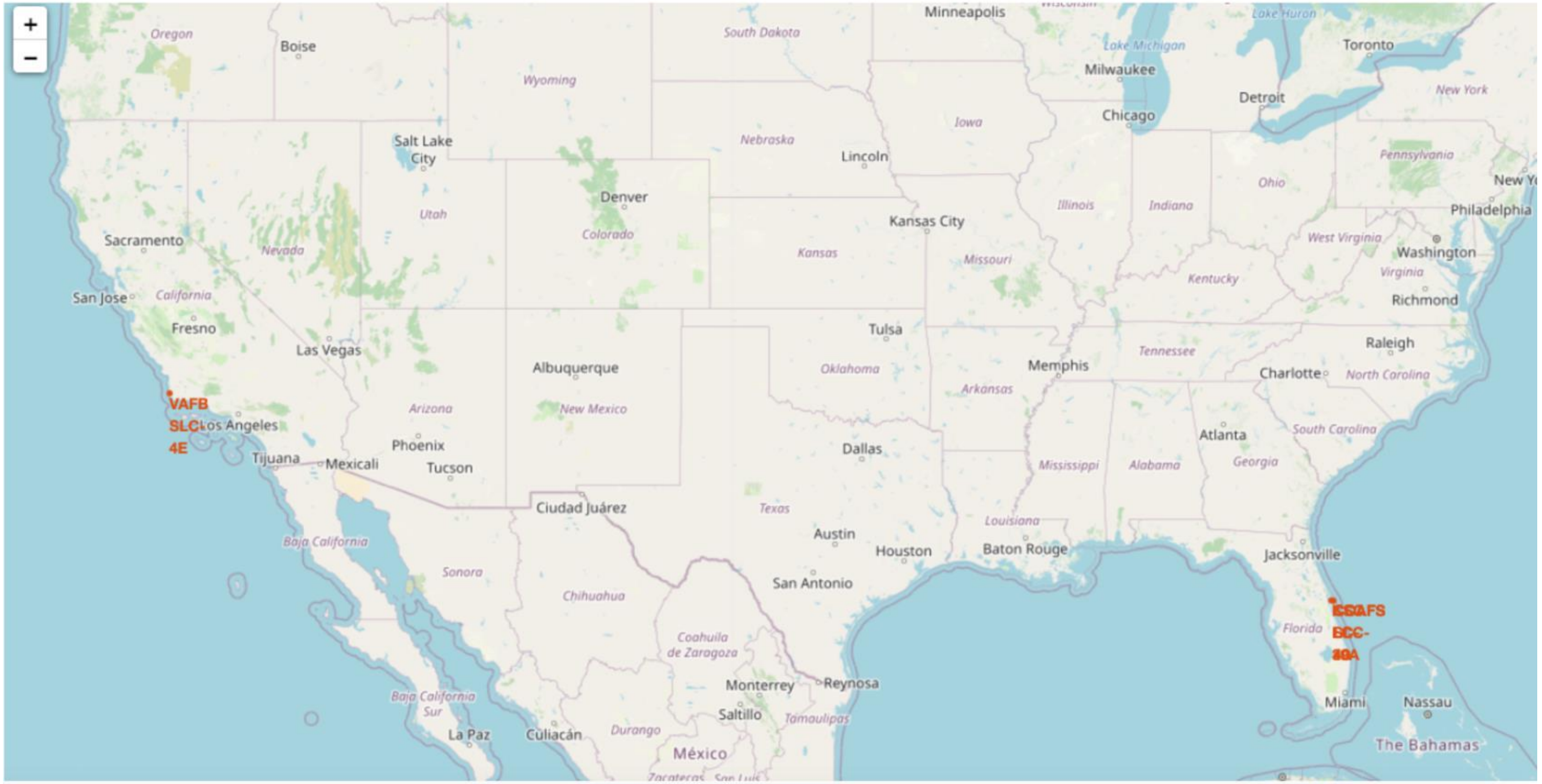
- The launch success yearly trend



RESULTS

3 Folium

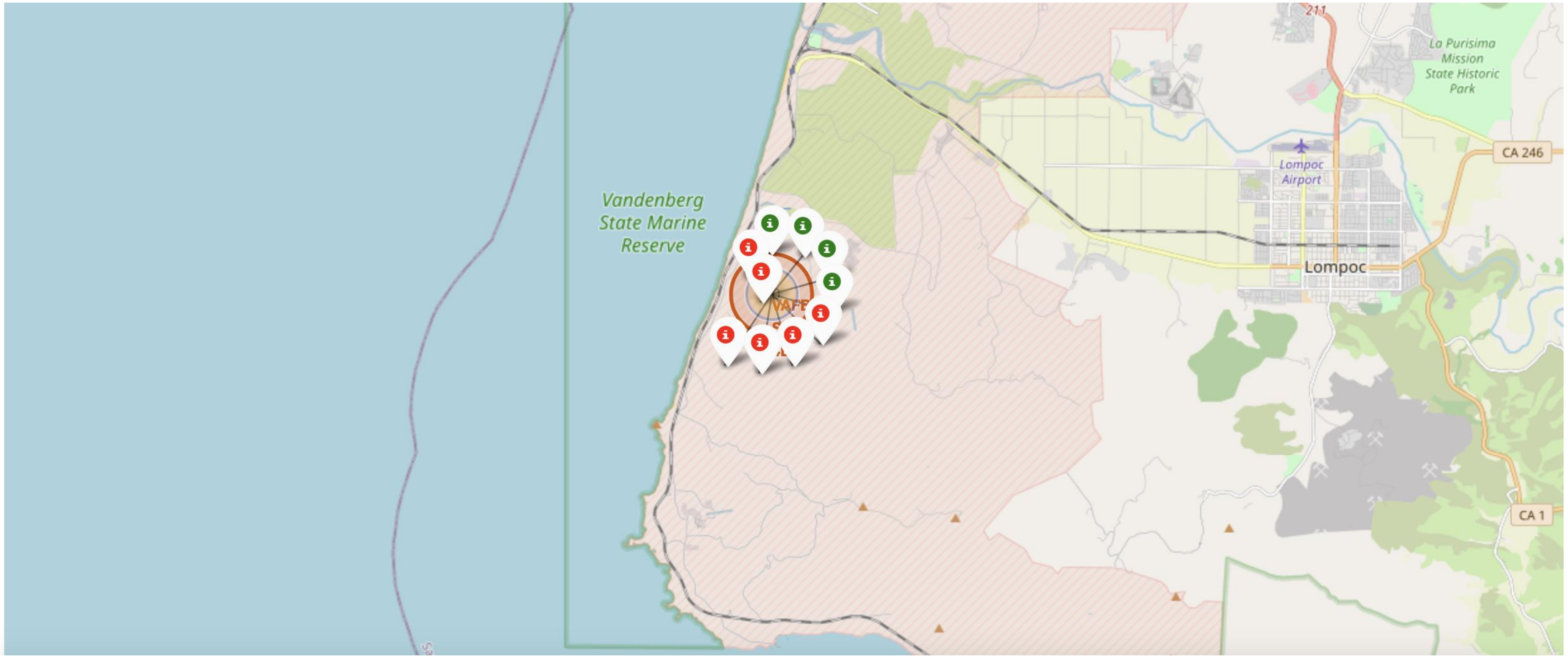
- All launch sites on map



RESULTS

3 Folium

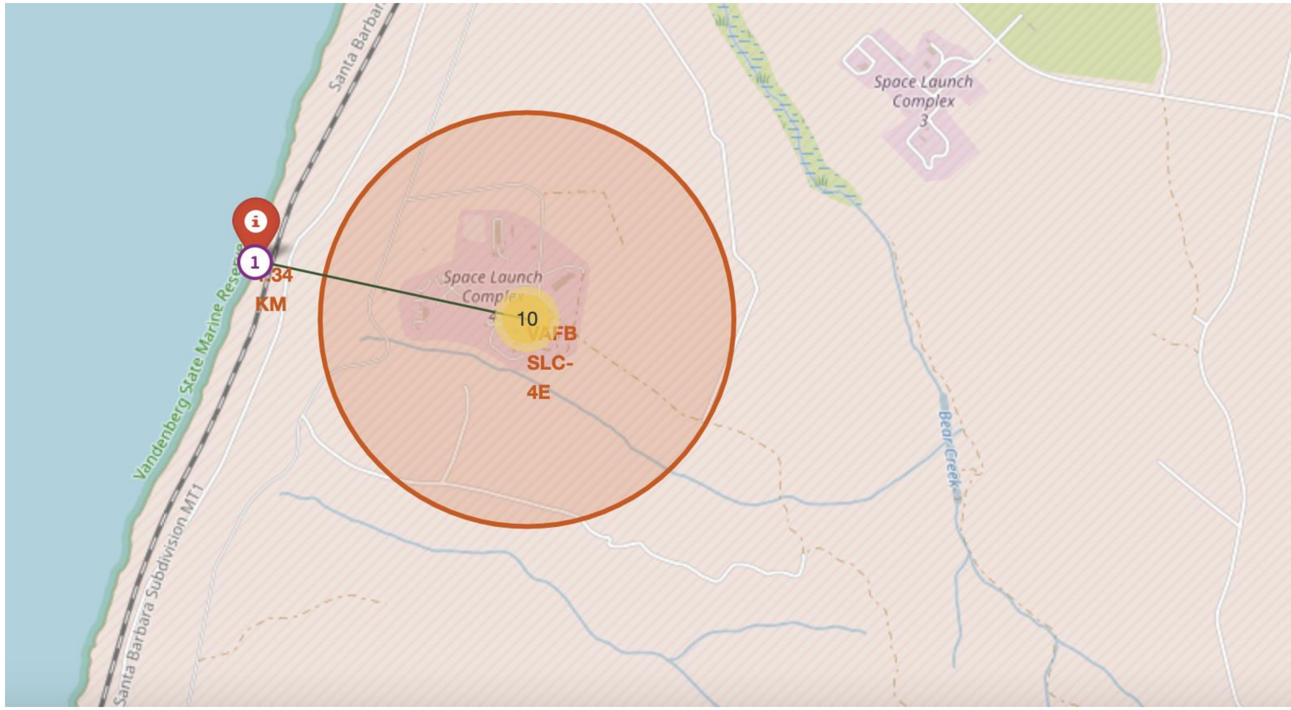
- The successful and unsuccessful launches for each site on the map
  - We can see green and red markers if we zoom in on one of the launch sites. Each green tag indicates a successful launch, whereas each red tag indicates a failure launch.



RESULTS

3 Folium

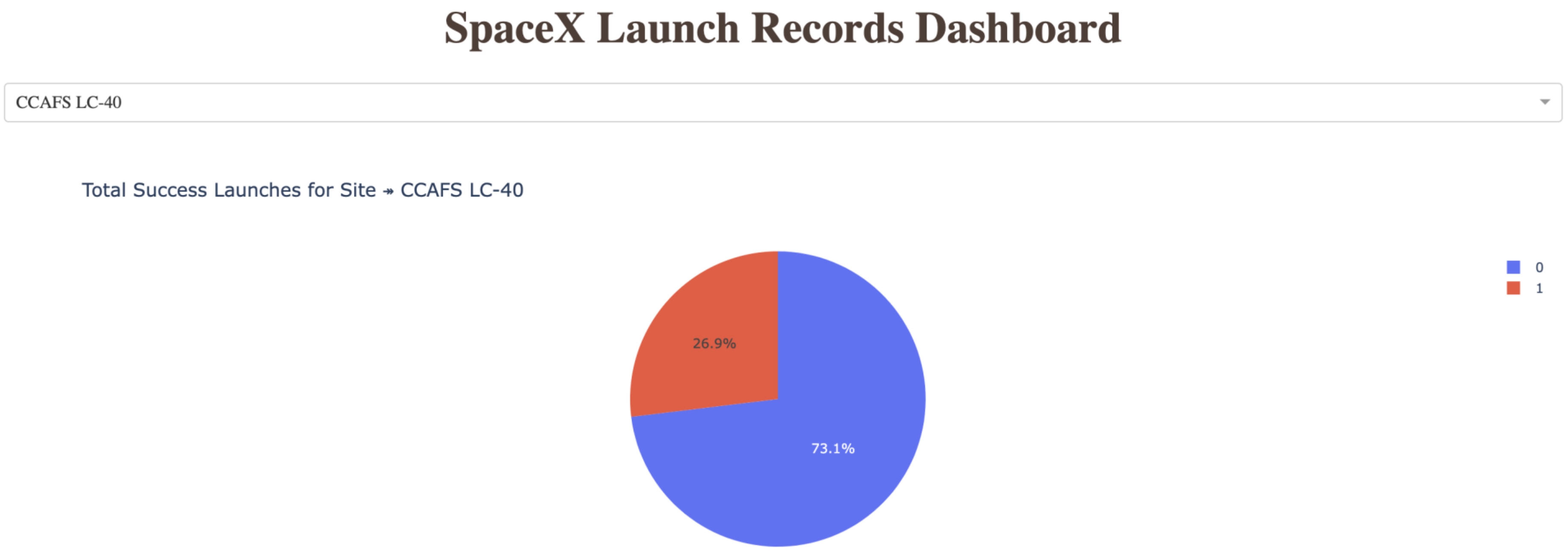
- Distances between a launch location to its surroundings, such as the next city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



RESULTS

4 Dash

- The image below depicts a pie chart when the launch site CCAFS LC-40 is selected..
- The number 0 denotes failed launches, whereas the number 1 symbolizes successful launches. We can see that 73.1% of the launches at CCAFS LC-40 are failures.



RESULTS

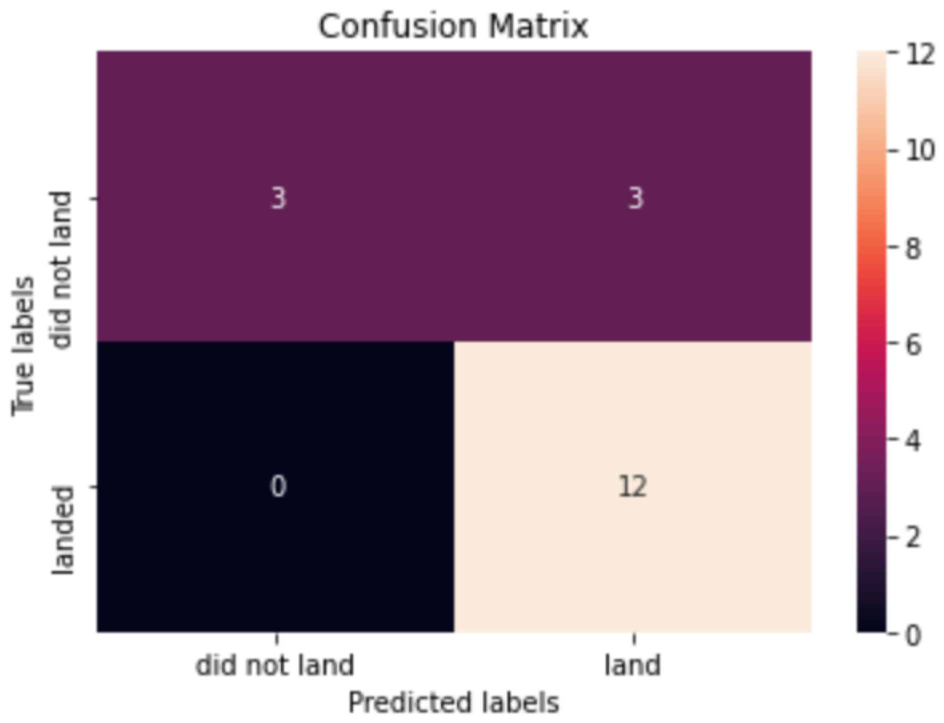
4 Dash

- The picture below depicts a scatterplot with a payload mass range of 2000kg to 8000kg.
- Class 0 denotes failed launches, and class 1 denotes successful launches.

RESULTS

5 Predictive Analysis

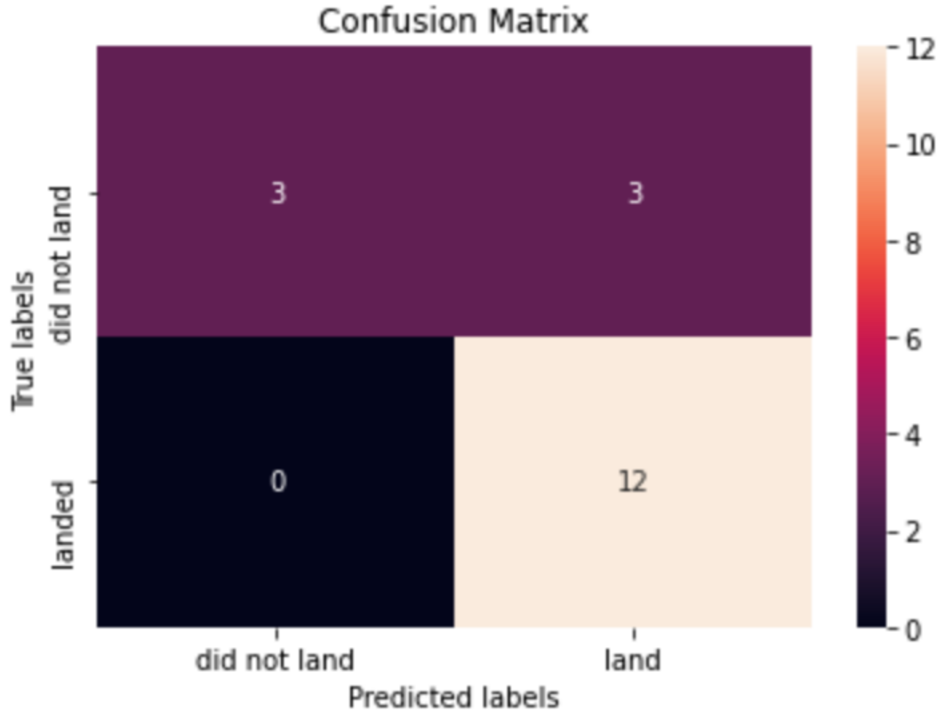
- Logistic regression
  - GridSearchCV best score: 0.8464285714285713
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

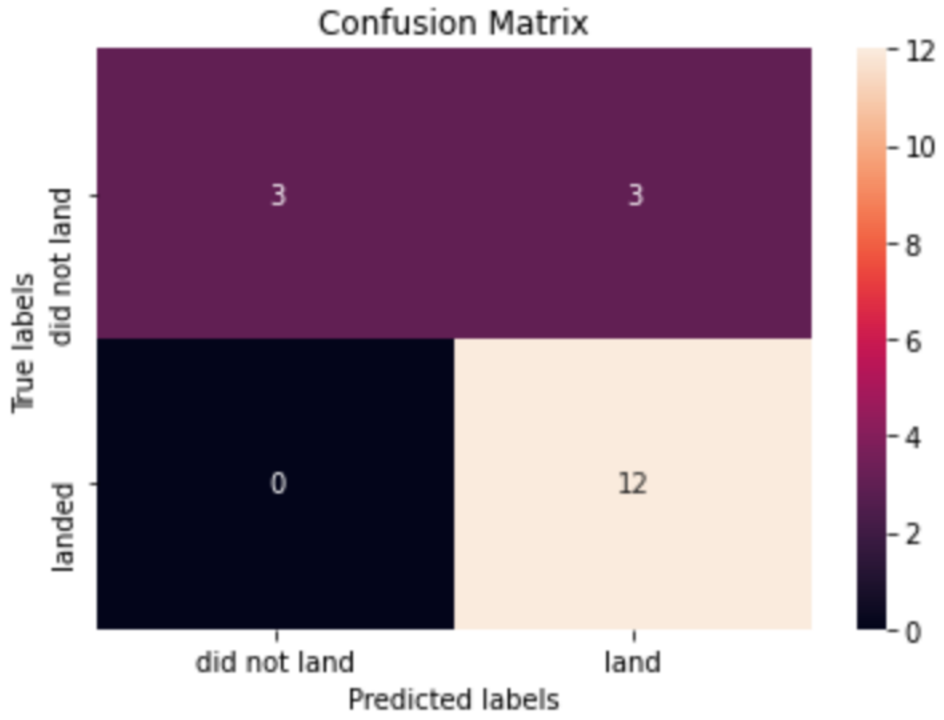
- Support vector machine (SVM)
  - GridSearchCV best score: 0.8482142857142856
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



RESULTS

5 Predictive Analysis

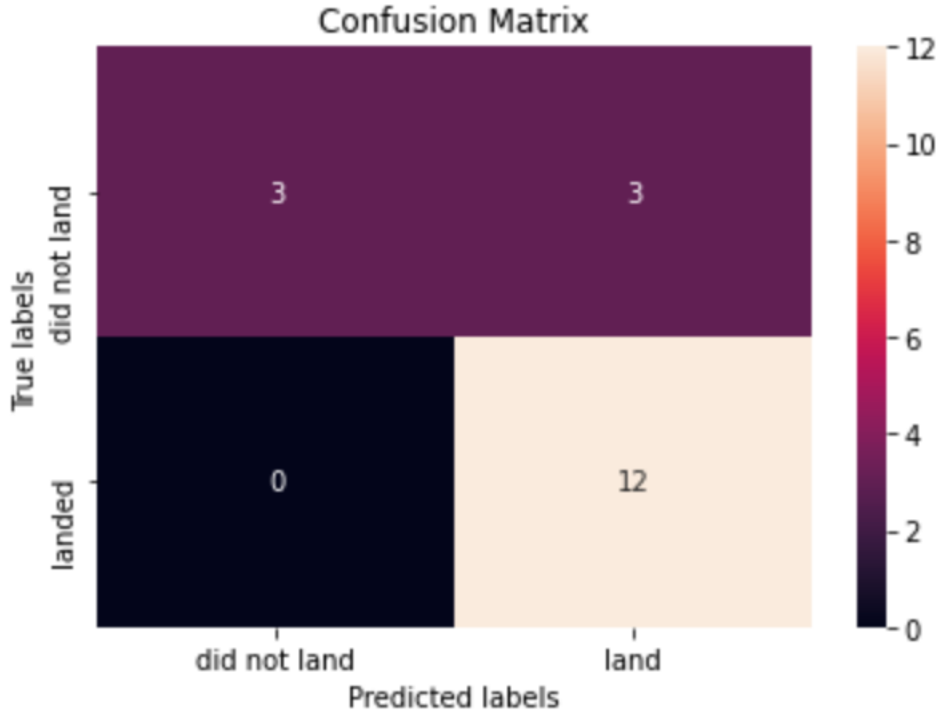
- Decision tree
  - GridSearchCV best score: 0.8892857142857142
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



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

- When we compare the results of all four models, we can observe that they all have the same accuracy score and confusion matrix when tested on the test set.
- As a result, their best GridSearchCV ratings are utilized to rank them instead. The models are rated in the following order based on their GridSearchCV best scores, with the first being the best and the last 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

- We can see from the data visualization section that some features may have an association with the mission outcome in a variety of ways. For example, with heavy payloads, the successful or positive landing rate is higher in orbit types such as Polar, LEO, and ISS. However, for GTO, we can't tell the difference because both positive and negative landing rates (missed missions) are present.
- As a result, each characteristic may have an effect on the final mission outcome. It is difficult to determine how each of these features affects the mission outcome. However, we can apply machine learning algorithms to learn the pattern of previous data and predict whether or not a mission would be successful based on the supplied features.

CONCLUSION

- In this capstone, we try to predict whether or not the first stage of a specific Falcon 9 launch will land in order to calculate the cost of the launch.
- Each feature of a Falcon 9 launch, such as payload mass or orbit type, may have an impact on the mission's result.
- Several machine learning techniques are employed to study the patterns in previous Falcon 9 launch data in order to create prediction models that can be used to predict the outcome of a Falcon 9 launch.
- Among the four machine learning algorithms used, the predictive model built by the decision tree method performed the best.