Applied Data Science Capstone

Baran Ağırbaş

May 7,2023

OUTLINE

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

Executive Summary

- In this latest project, it will be predicted whether the first stage of SpaceX Falcon 9 will land successfully using several machine learning classification algorithms.
- The main steps in this project include:
  - Data collection, discussion and formatting
  - Exploratory data analysis
  - Interactive data visualization
  - Machine learning prediction
- The graphs show that some features of rocket launches are related to the outcome of the launches, namely success or failure.
- It was also concluded that the decision tree may be the best machine learning algorithm to predict whether the first stage of Falcon 9 will land successfully.

Introduction

- This model will predict whether the first stage of Falcon 9 will land successfully. SpaceX advertises Falcon 9 rocket launches costing $62 million on its website; other providers cost more than $165 million each, most of the savings being SpaceX's ability to reuse the first stage. Therefore, if it can be determined whether the first stage will land or not, the cost of the launch can also be determined. This information can be used if an alternative company wants to bid against SpaceX for a rocket launch.

- Most failed landings are planned, and sometimes SpaceX makes a controlled landing in the ocean.

- The main question that is being tried to be answered is the payload mass, trajectory type, launch location, etc. associated with a Falcon 9 rocket launch. For a given set of features including, will the first stage of the rocket land successfully?

Methodology

- The overall methodology includes:
  - Data collection, wrangling, and formatting, using:
    - SpaceX API
    - Web scraping
  - Exploratory data analysis (EDA), using:
    - Pandas and NumPy
    - SQL
  - Data visualization, using:
    - Matplotlib and Seaborn
    - Folium
    - Dash
  - Machine learning prediction, using
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Decision Tree
    - K-nearest Neighbors (KNN)

Methodology

- 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.

Methodology

- 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.

Methodology

- 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

- 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

- 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

- 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

- 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 - SQL

- The names of the unique launch sites in the space mission

```
%%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL
 * sqlite:///my_data1.db
Done.
```

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

- 5 records where launch sites begin with ‘CCA’

```
%%sql SELECT * FROM SPACEXTBL 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
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Results - SQL

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

```
[9]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total payload mass by NASA (CRS)" FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
 * sqlite:///my_data1.db
Done.
[9]:
```

Total payload mass by NASA (CRS)
45596

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

```
[10]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average payload mass by Booster Version F9 v1.1" FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
 * sqlite:///my_data1.db
Done.
[10]:
```

Average payload mass by Booster Version F9 v1.1
2928.4

- 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 - SQL

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

- The total number of successful and failure mission outcomes

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

number_of_success_outcomes	number_of_failure_outcomes
100	1

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

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

Results - 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

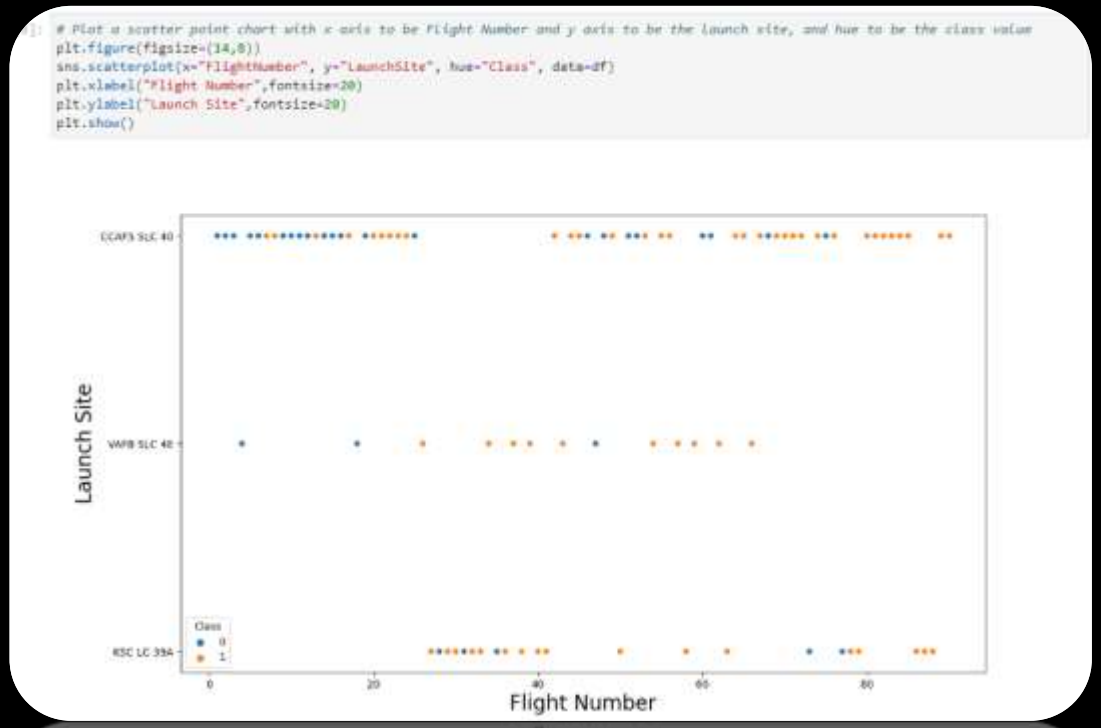
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

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 – Matplotlib and Seaborn

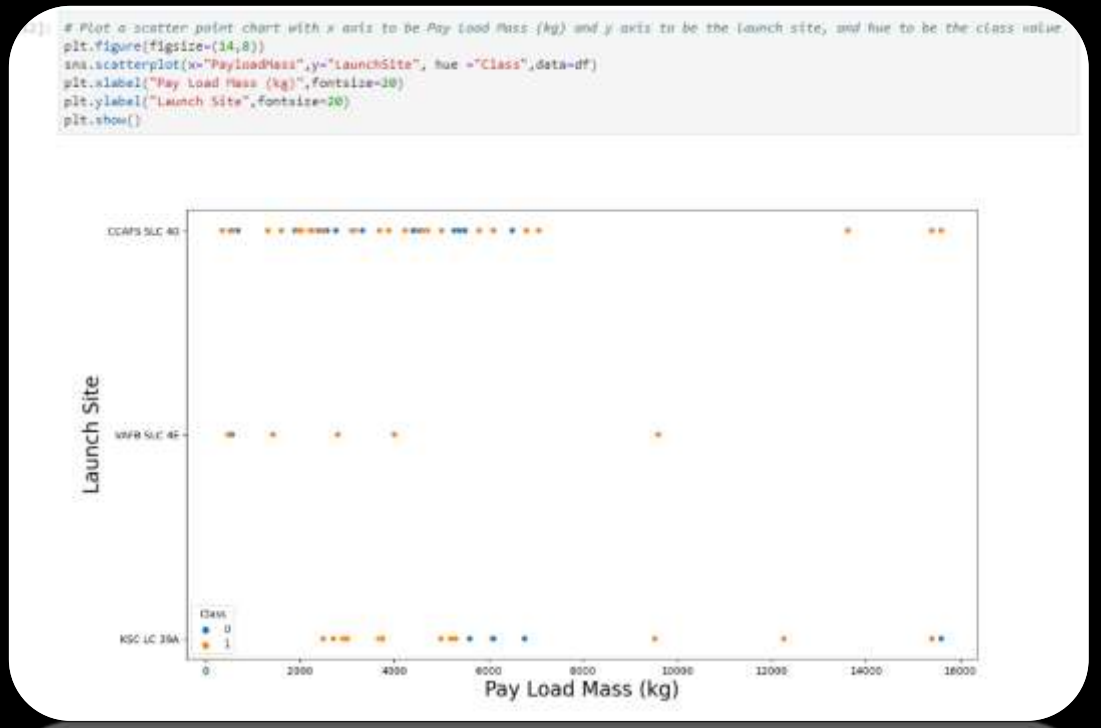
- The relationship between flight number and launch site

```
# Plot a scatter point chart with x axis to be Flight Number and y axis to be the Launch site, and hue to be the class value
plt.figure(figsize=(14,8))
sns.scatterplot(x="FlightNumber", y="LaunchSite", hue="Class", data=df)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```



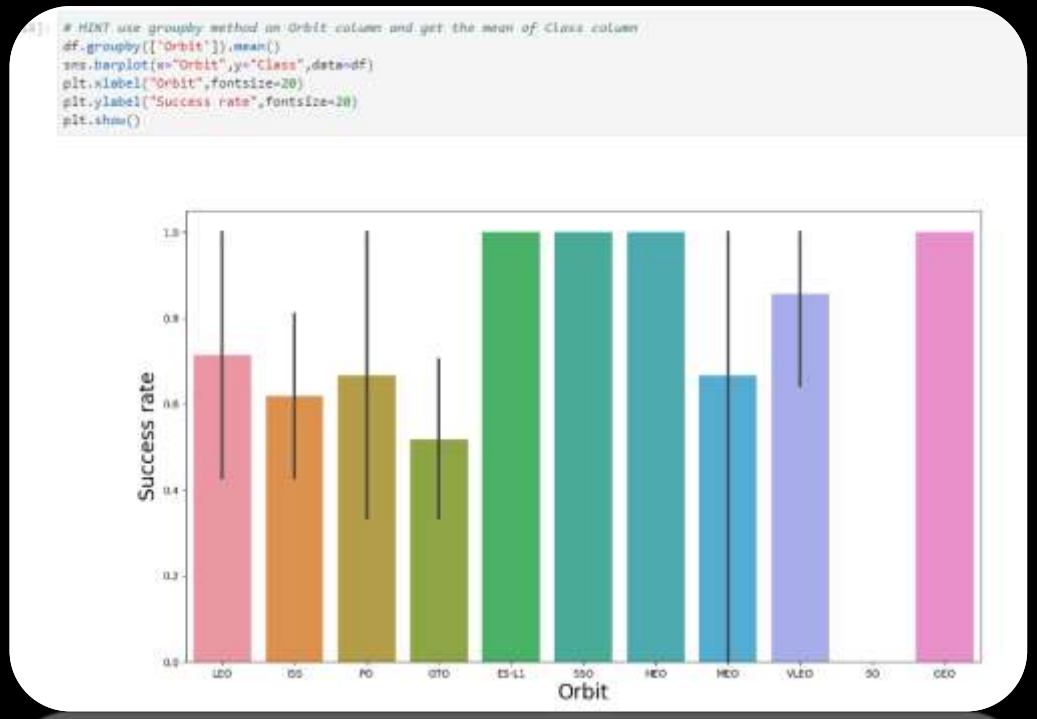
- The relationship between payload mass and launch site

```
# Plot a scatter point chart with x axis to be Pay Load Mass (kg) and y axis to be the Launch site, and hue to be the class value
plt.figure(figsize=(14,8))
sns.scatterplot(x="PayloadMass",y="LaunchSite", hue ="Class",data=df)
plt.xlabel("Pay Load Mass (kg)",fontsize=20)
plt.ylabel("Launch Site",fontsize=20)
plt.show()
```

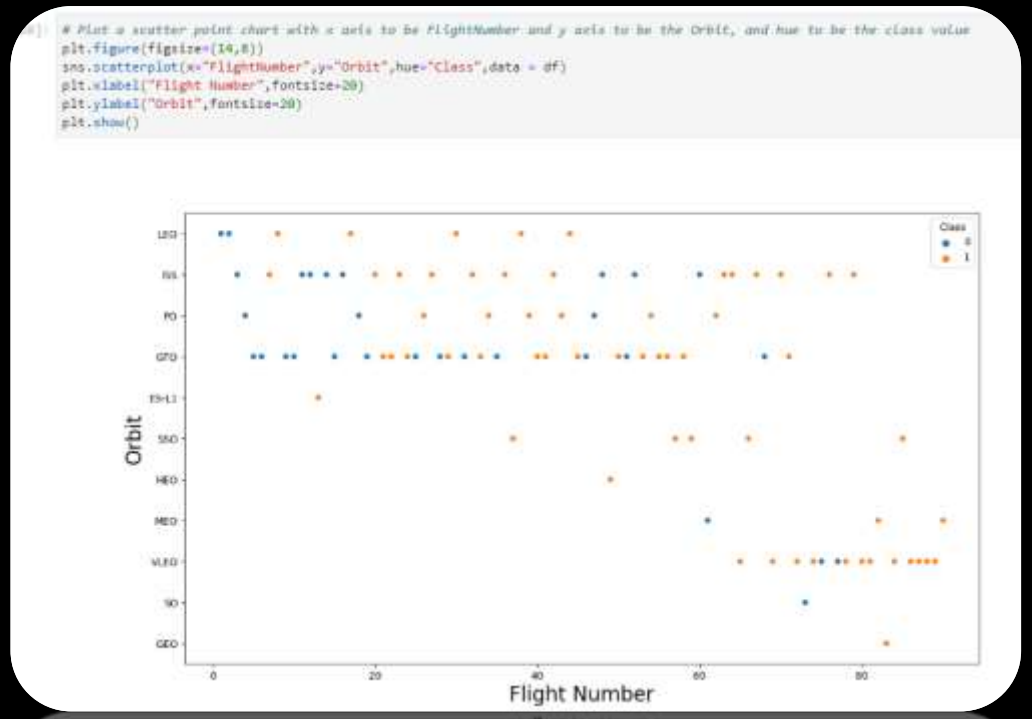


Results – Matplotlib and Seaborn

- The relationship between success rate and orbit type

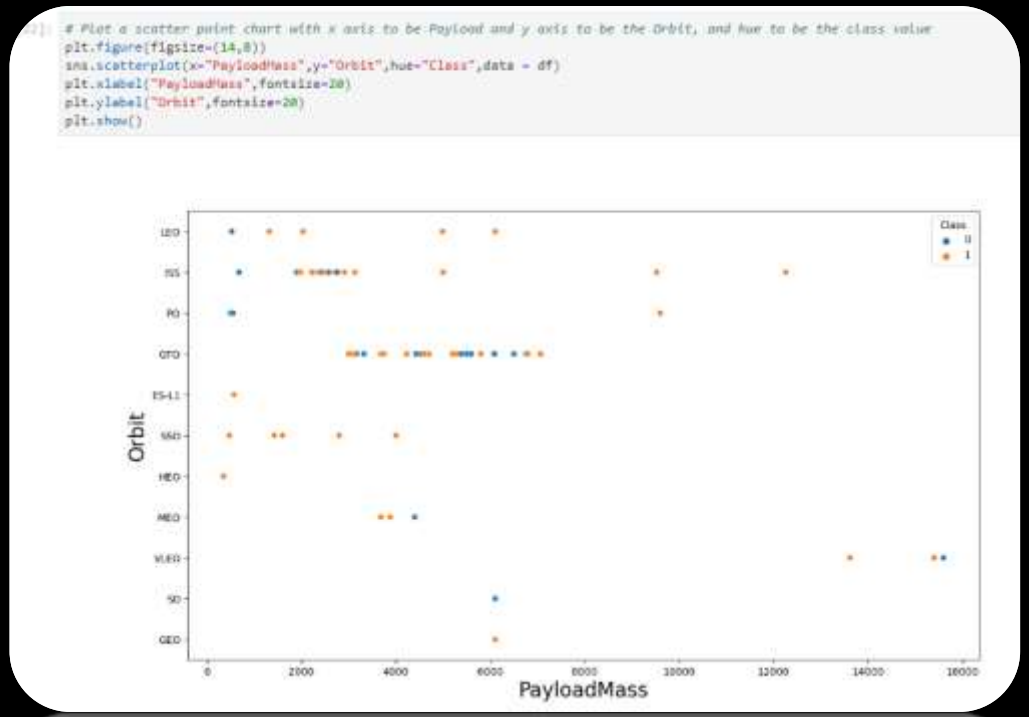


- The relationship between flight number and orbit type

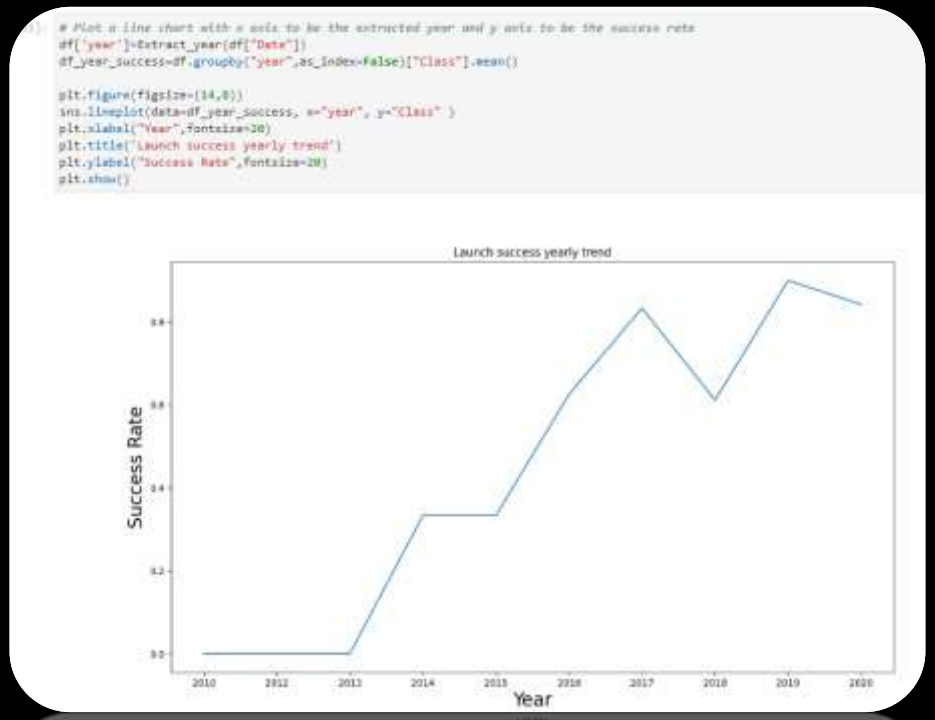


Results – Matplotlib and Seaborn

- The relationship between payload mass and orbit type



- The launch success yearly trend

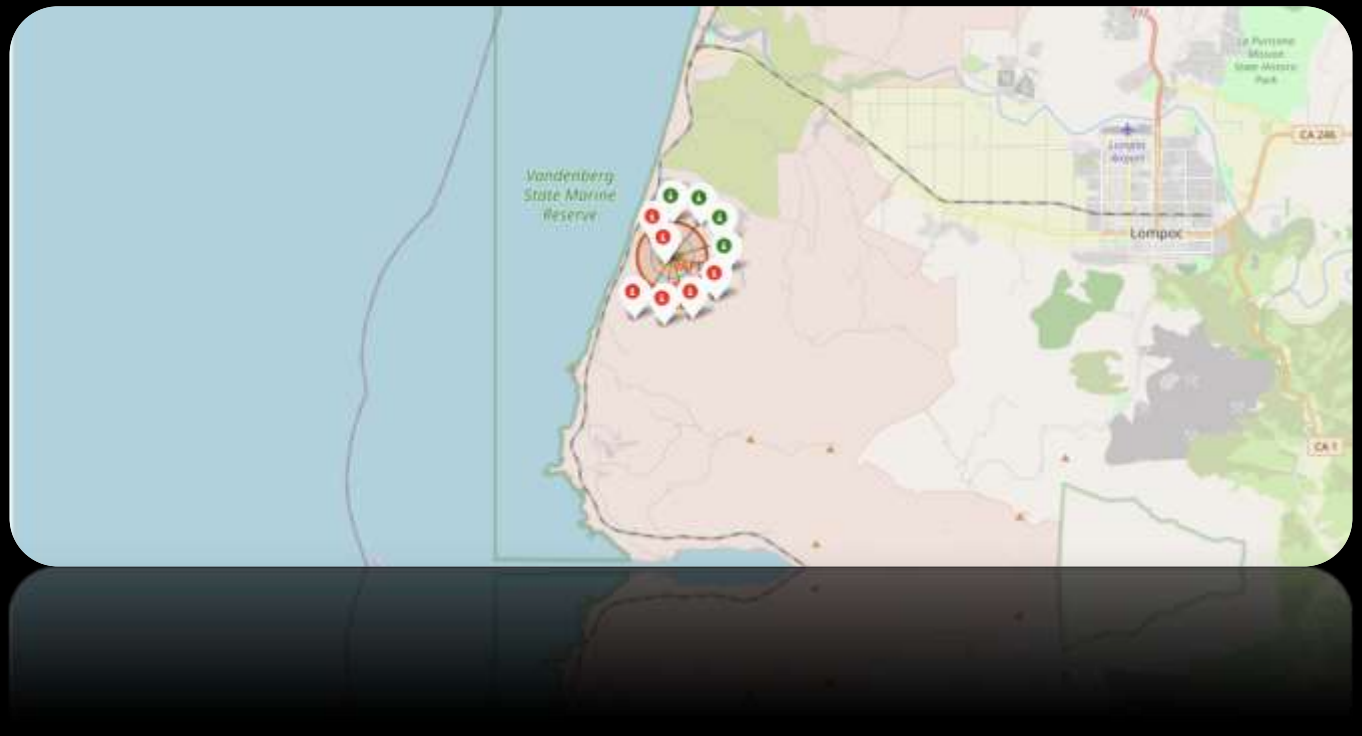


Results – Folium

- All launch sites on map

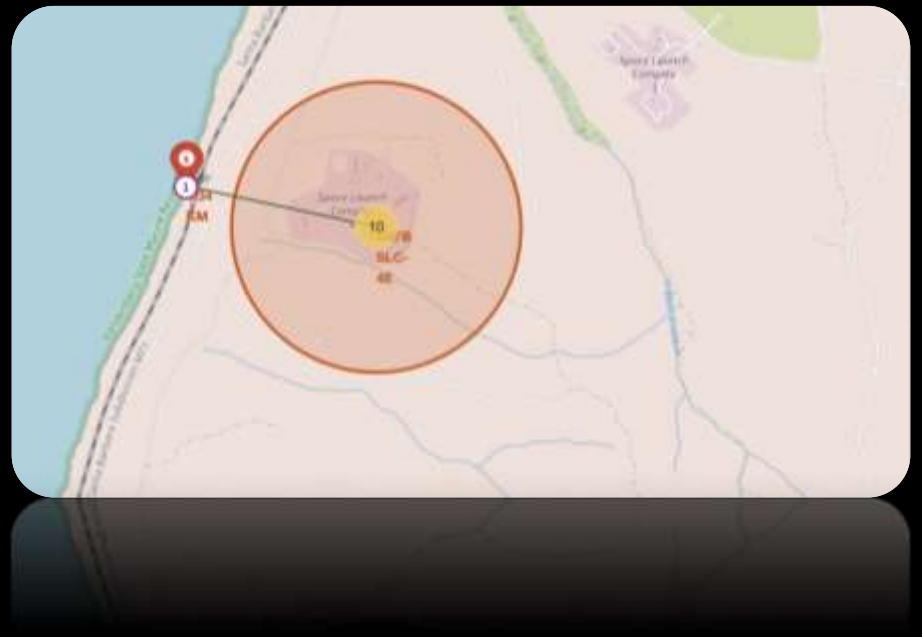
Results – Folium

- 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



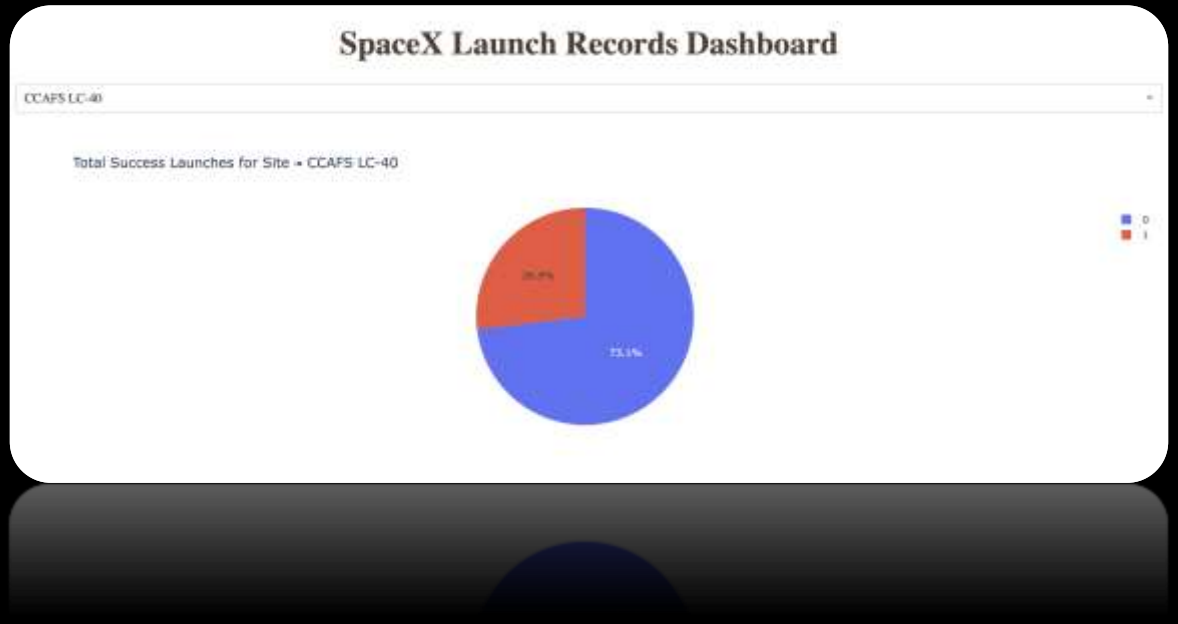
Results – Folium

- 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 – 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.

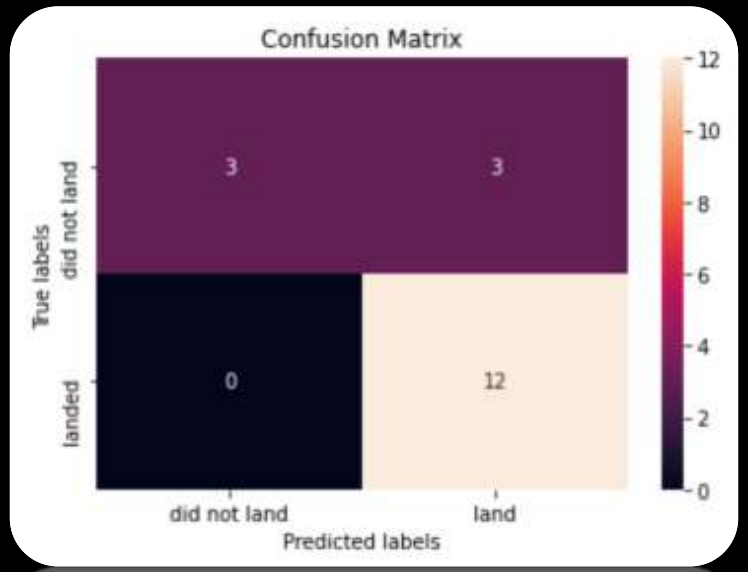


Results – Dash

- 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.

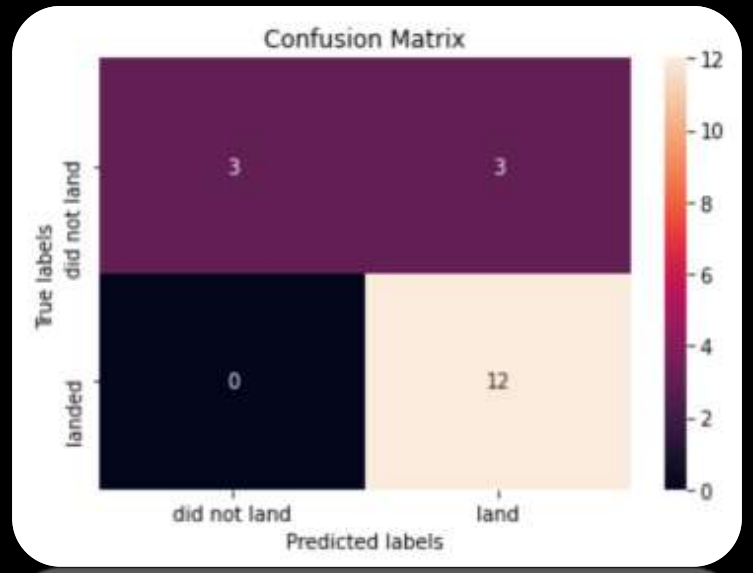
Results – Predictive Analysis

- Logistic regression
  - GridSearchCV best score: 0.8465
  - Accuracy score on test set: 0.8334



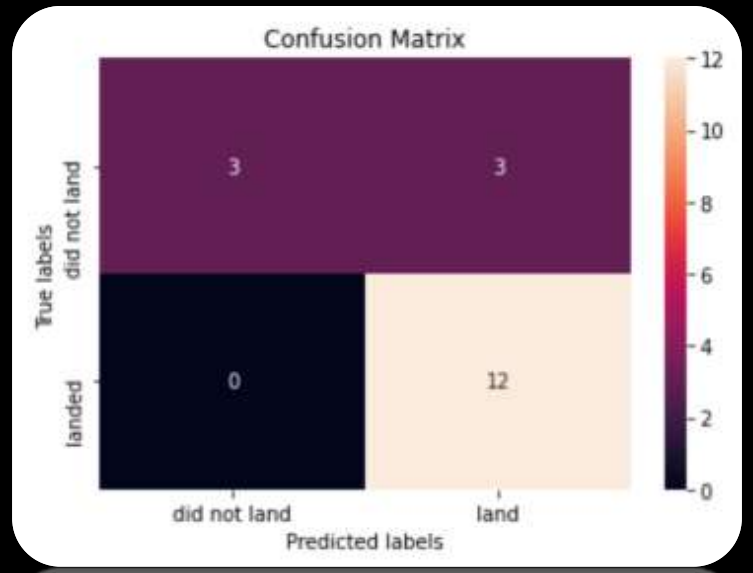
Results – Predictive Analysis

- Support vector machine (SVM)
  - GridSearchCV best score: 0.8483
  - Accuracy score on test set: 0.8334



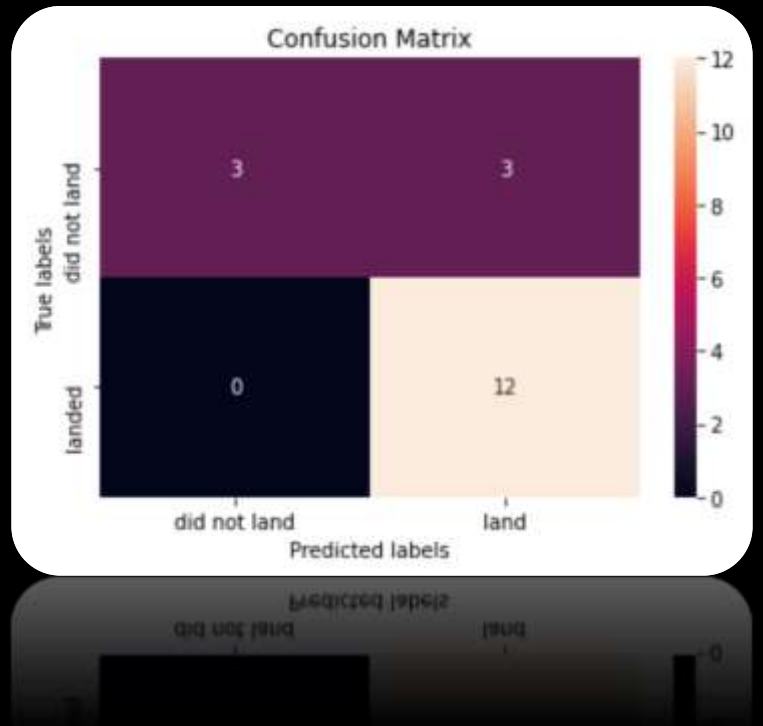
Results – Predictive Analysis

- Decision tree
  - GridSearchCV best score: 0.8893
  - Accuracy score on test set: 0.8334



Results – Predictive Analysis

- K nearest neighbors (KNN)
  - GridSearchCV best score: 0.8483
  - Accuracy score on test set: 0.8334



Results – 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: 0.8892857)
    2. K nearest neighbors, KNN (GridSearchCV best score: 0.8482142)
    3. Support vector machine, SVM (GridSearchCV best score: 0.8482142)
    4. Logistic regression (GridSearchCV best score: 0.8464285)

Discussion

- From the data visualization section, it can be seen that some properties can be related to the task result in various ways.
- For example, the rate of successful landing or positive landing with heavy payloads is greater for Polar, LEO and ISS orbit types. But for the GTO, it is not well distinguishable as both positive landing rate and negative landing are here.
- Therefore, each feature can have a certain impact on the final task outcome. It is difficult to come up with precise ways of how each of these characteristics affects the task outcome.
- However, some machine learning algorithms can be used to learn the pattern of historical data and predict whether a task will be successful based on the given features.

Results – Predictive Analysis

- In this project, an attempt was made to predict whether the first stage of a given Falcon 9 launch would land in order to determine the cost of a launch.
- Each aspect of a Falcon 9 launch, such as payload mass or orbit type, can affect mission outcome in a particular way.
- Various machine learning algorithms are used to learn patterns from past Falcon 9 launch data to produce predictive models that can be used to predict the outcome of a Falcon 9 launch.
- The prediction model produced by the decision tree algorithm showed the best performance among the 4 machine learning algorithms used.