APPLIED DATA SCIENCE CAPSTONE

SAI DURGA VANNALA

AUGUST 4, 2023.

Space x Falcon 9 First Stage Landing prediction

SAI DURGA VANNALA.

OUTLINE

- EXECUTIVE SUMMARY
- INTRODUCTION
- METHODOLOGY
- RESULTS
- DISCUSSIONS
- CONCLUSIONS

Contoso

EXECUTIVE SUMMARY

- This project will predict whether the Falcon 9 first stage will land successfully using machine learning classification algorithms.
- The steps involved in this project are:
    1. Data collection
    2. Exploratory Data Analysis(EDA)
    3. Data visualization
    4. Machine Learning Prediction
- Using visualization it has shown the aspects of rocket launches that are related to the outcome whether it is success or failure
- It concludes that decision tree is the best machine learning for predicting whether Falcon9 first stage will land successfully.

Contoso

INRODUCTION

- This project will predict whether the Falcon 9 first stage will land successfully. 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 characteristics such as cargo tonnage, orbit type, launch site, and so on, the first stage of the rocket will successfully land.

Contoso

METHODOLOGY

METHODOLOGY

Executive Summary

1. Data Collection and data wrangling.

- Space X API
- Web Scraping

2. Exploratory Data Analysis

- Pandas and Numpy
- SQL

3. Data Visualization

- Matplotlib and Seaborn
- Folium
- Dash

4. Machine Learning Prediction

- Logistic Regression
- SVM
- Decision Tree
- K-Nearest Neighbors(KNN)

Data Collection

- Data collection is the process of acquiring and measuring information on certain variables in an established system, allowing one to answer pertinent questions and evaluate outcomes. As previously stated, the dataset was obtained from Wikipedia via REST API and web scraping.

- For REST API, begins with a get request. The response content was then encoded as JSON and converted into a pandas data frame using json_normalize(). We then cleaned the data, looked for missing numbers, and filled in the gaps.

- BeautifulSoup is used for web scraping to extract the launch records as an HTML table, parse the table, and convert it to a pandas dataframe for further analysis.

Contoso

Data collection-Space X API

- Request API for space X launch data using the API provided.
- The data is filtered to just include Falcon 9 launches because the API contains information about a variety of rocket launches carried out by SpaceX.
- Each missing value is replaced by the mean of the column to which it belongs in the data.
- The few rows and columns of the data is shown in the image.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None

Data collection-Web Scraping

- Performed web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled List of Falcon 9 and Falcon Heavy launches.
- Requested the Falcon 9 launch wiki page from its URL.
- Using BeautifulSoup extracted Falcon 9 launch records from Wikipedia and parsed the table and converted them into a pandas data frame.

Contoso

Data Wrangling

- Performed exploratory data analysis and determined the training labels.
- Calculated the number of launces on each site.
- Calculated the number and occurrence of mission outcome of the orbits.
- Created a landing outcome label from outcome column.

Contoso

Exploratory Data Analysis

- **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 occurrences 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 they are as follows.
  - 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

Contoso

EDA with Visualization

Matplotlib and Seaborn

- Scatterplots, bar charts, and line charts are utilized to visualize the data using functions from the Matplotlib and Seaborn libraries.
- Plots and charts are employed to learn more about the connections between various features, including:
  - The correlation between the launch site and the flight number; the correlation between the bulk of the cargo and the launch site; and the correlation between the success rate and the orbit type.

Folium

- Interactive maps are used to display the data using functions from the Folium libraries.
  - Marked all launch sites on a map
  - Marked the success/failed launches for each site on the map
  - Calculated the distances between a launch site to its proximities.

Dash

- In order to create an interactive site where we may switch the input using a dropdown menu and a range slider, Dash functions were implemented.
- Using a pie chart and a scatterplot, the interactive site shows:
  - The total success of launches from each launch site
  - The correlation between payload mass and mission outcome (success or failure) for each launch site.

Machine Learning Prediction

- Machine learning models are built using functions from the Scikit-learn toolkit.
- The steps in the machine learning prediction phase are as follows:
  - Standardising the data, separating it into training and test sets, and developing machine learning models, which comprise:
  - Logistic Regression
  - Support Vector machines
  - Decision Tree
  - K-Nearest Neigbors
- Models should be evaluated based on their accuracy scores and confusion matrices, and they should be fitted to the training set, with the optimal hyperparameter combination found for each model.

RESULTS

- EDA WITH SQL
- EDA WITH DATA VISUALIZATION
- DATA VISUALIZATION WITH FOLIUM
- DATA VISUALIZATION WITH DASH
- MACHINE LEARNING PREDICTION

Results-EDA with SQL

1. Unique launch site names

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

2. 5 records that start with string “CCA”.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Contoso

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

sum(PAYLOAD_MASS__KG_)
45596

4. Average payload mass carried by booster version F9 v1.1

avg(PAYLOAD_MASS__KG_)
2928.4

5. Date of the first successful landing outcome in ground pad was acheived.

min(DATE)
2015-12-22

6. 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

7. The total number of successful and failure mission outcomes.

count(MISSION_OUTCOME)
99

Contoso

8. The names of the booster_versions which have carried the maximum payload mass.

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

Contoso

9. 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

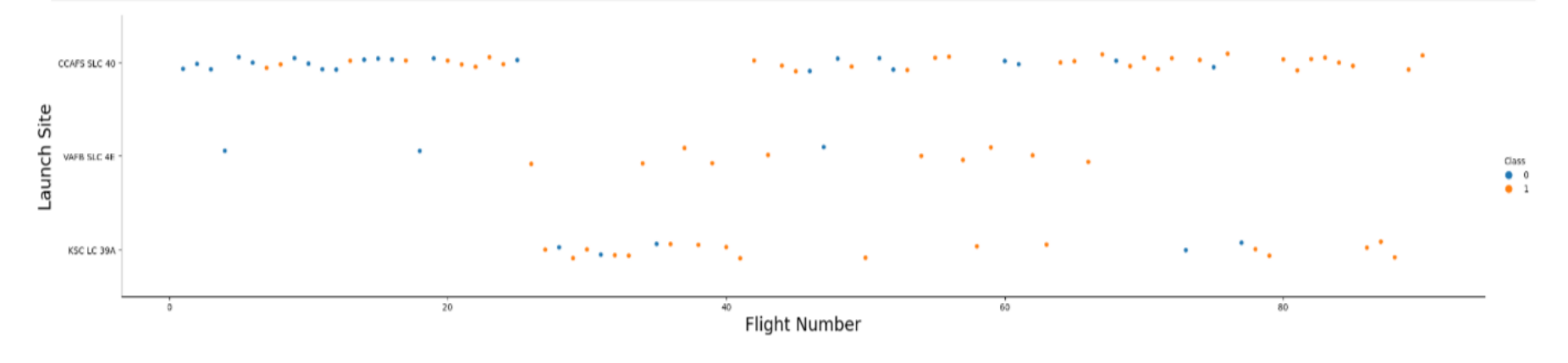
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

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

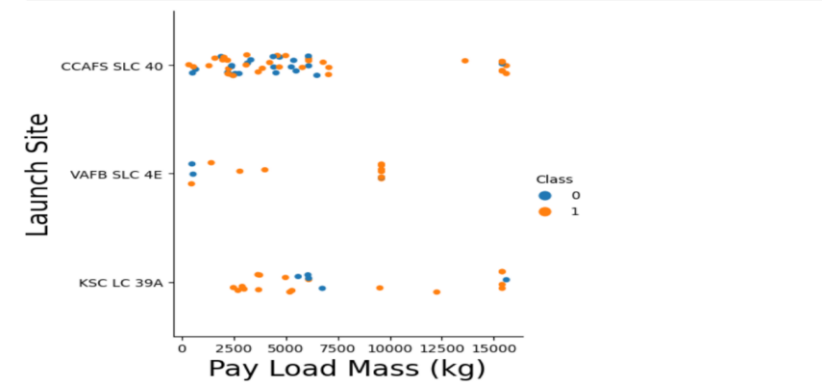
Contoso

Results-EDA with Visualization

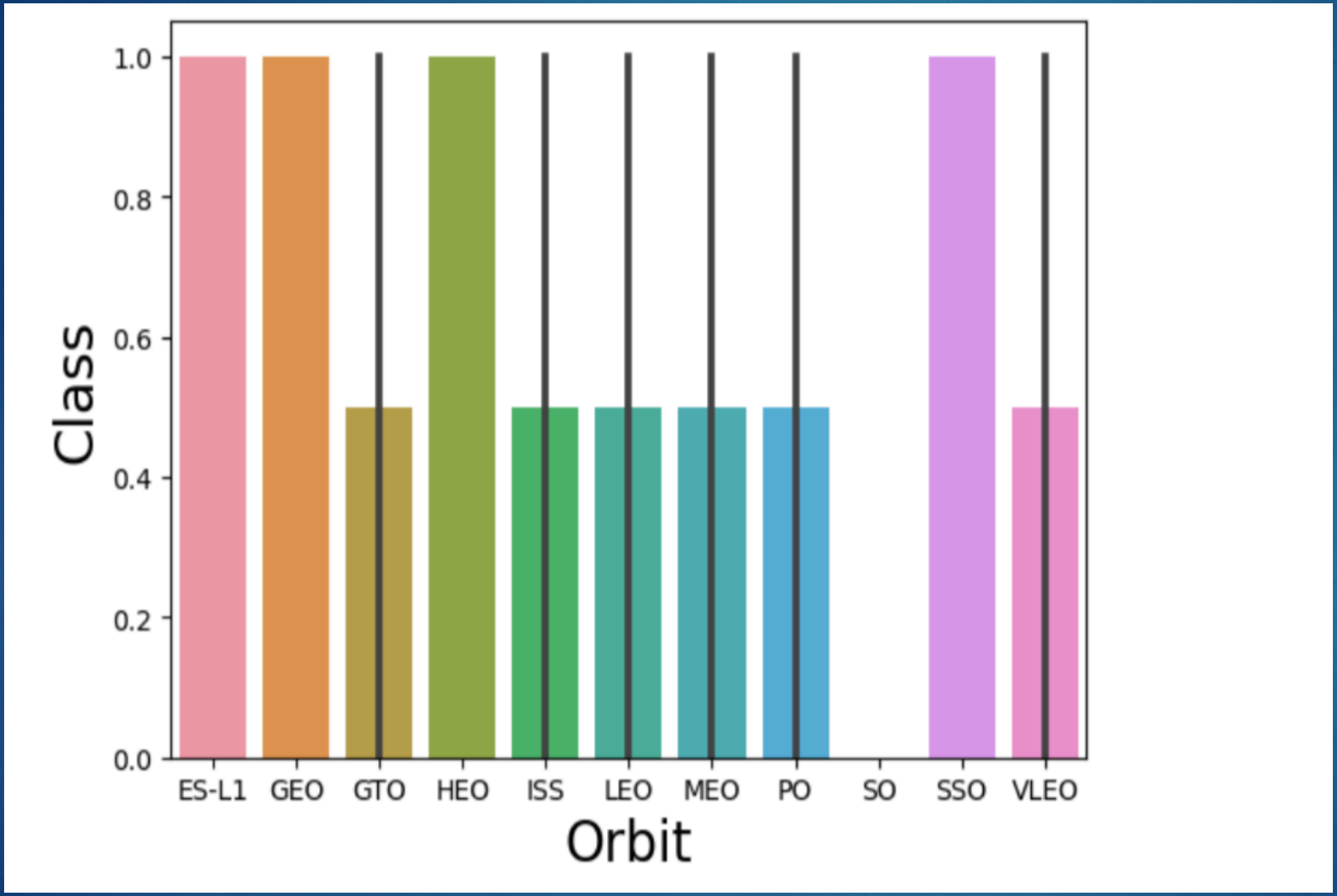
1. Relationship between Flight Number and Launch Site.



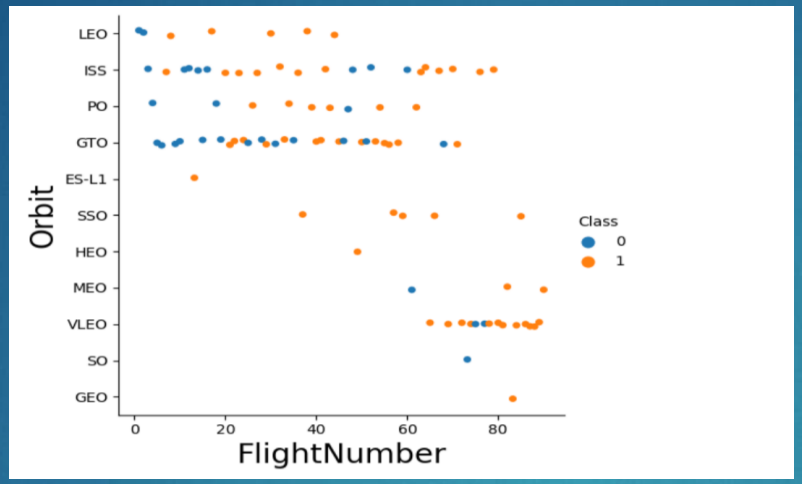
2. Relationship between Payload and Launch Site



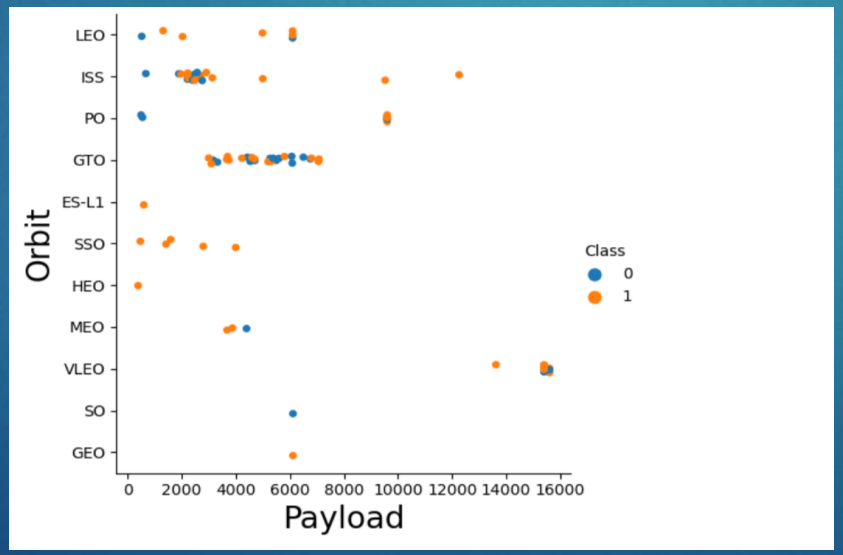
3. Relationship between success rate of each orbit type using bar chart



4. Relationship between Flight Number and Orbit type



5. Relationship between Payload and Orbit type

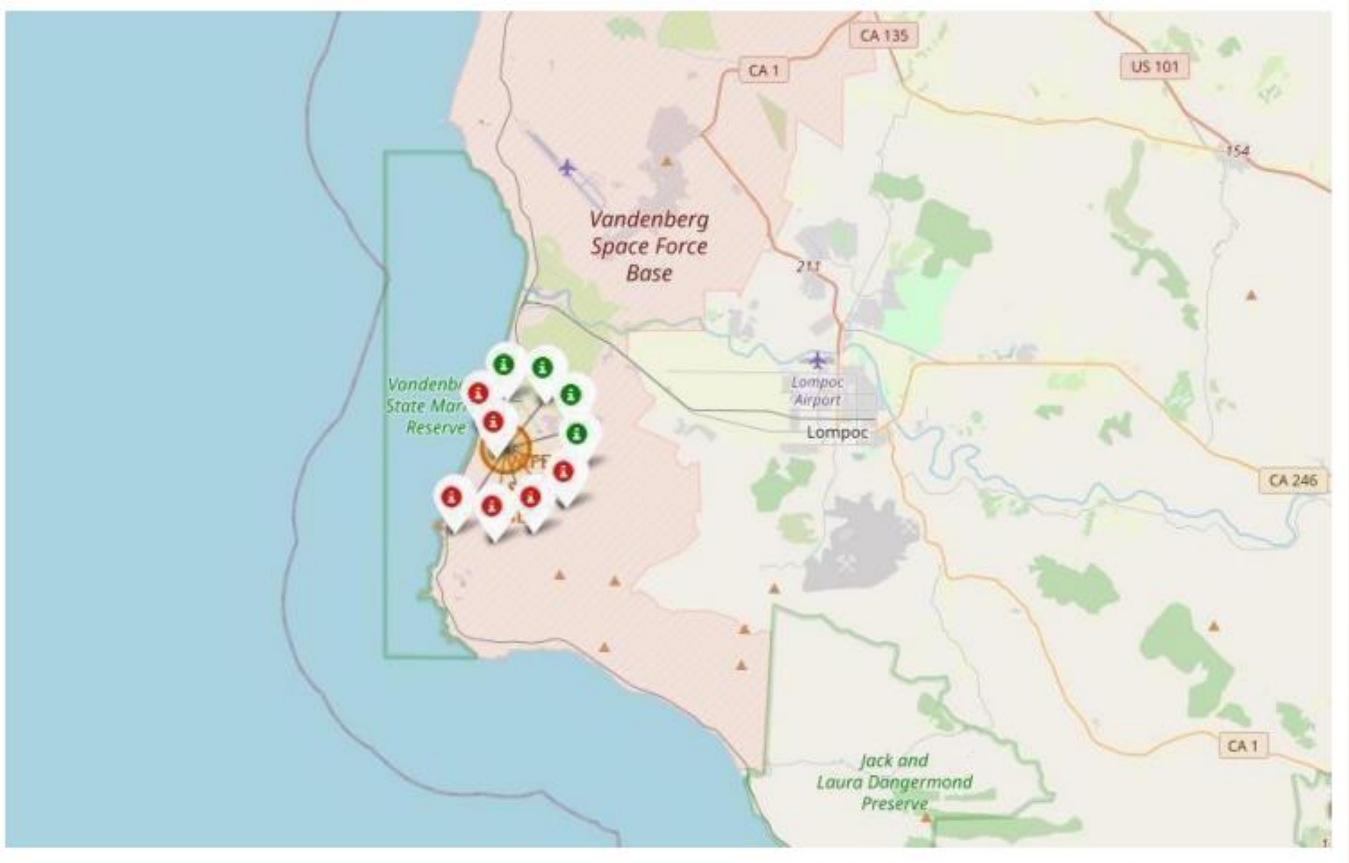


Contoso

Results- Data Visualization with Folium

1. Launch Site Locations

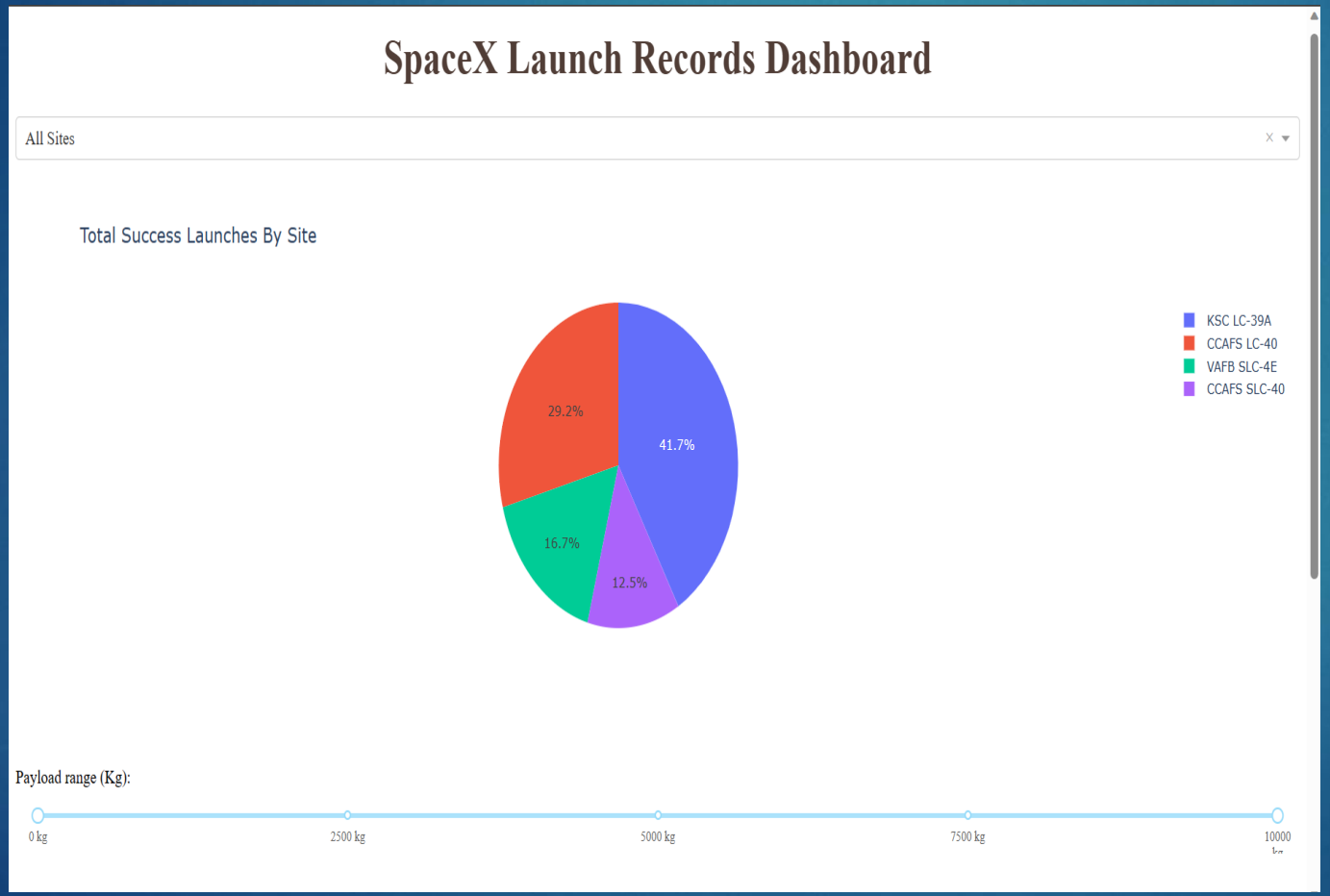
2. Success/failed launches for each site on the map



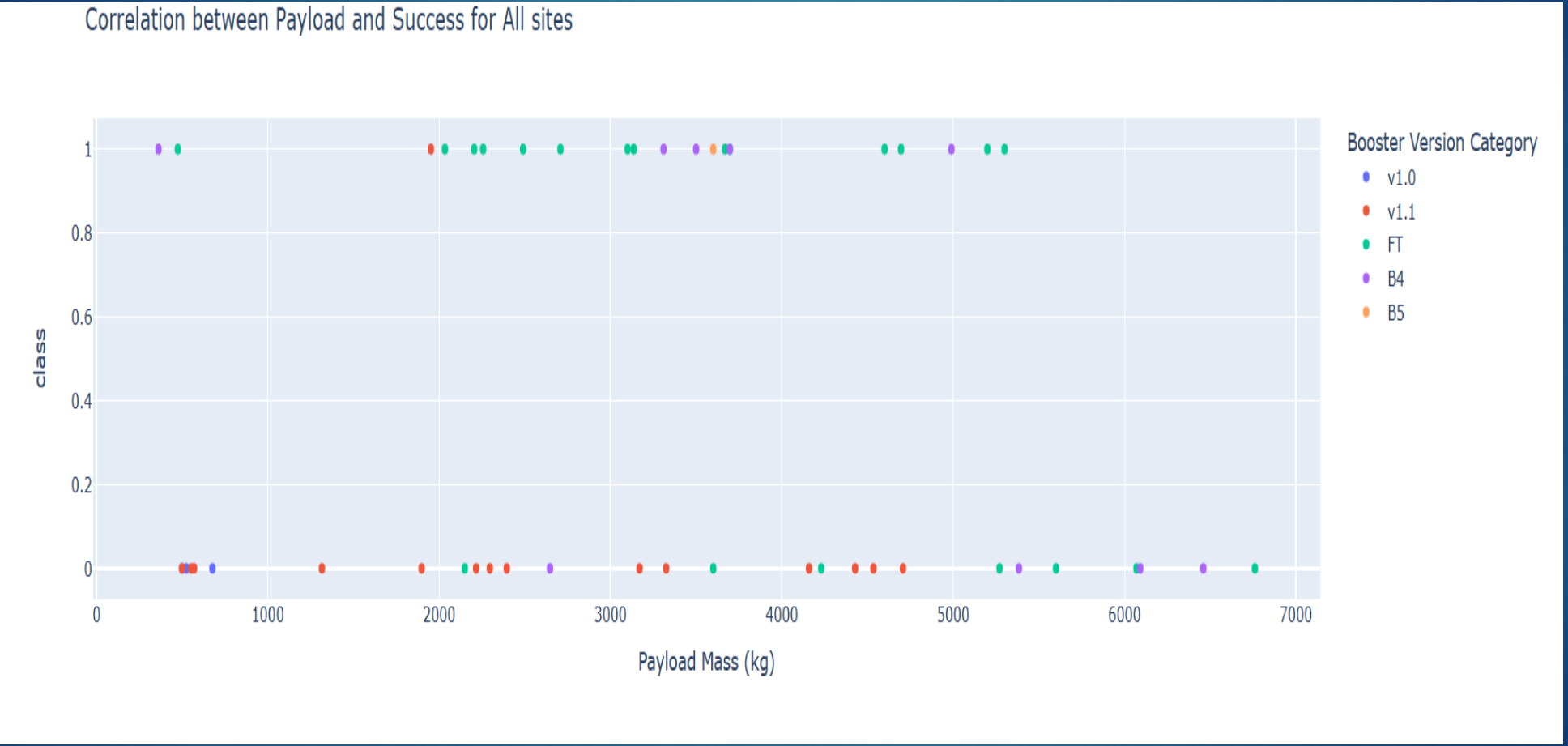
Green indicates successful launches and red indicates failed launches.

Results-Data Visualization with Dash

Pie Chart for successful launches across launch sites



Correlation between Payload and Success for All sites.



Results- Machine Learning Prediction

Logistic Regression

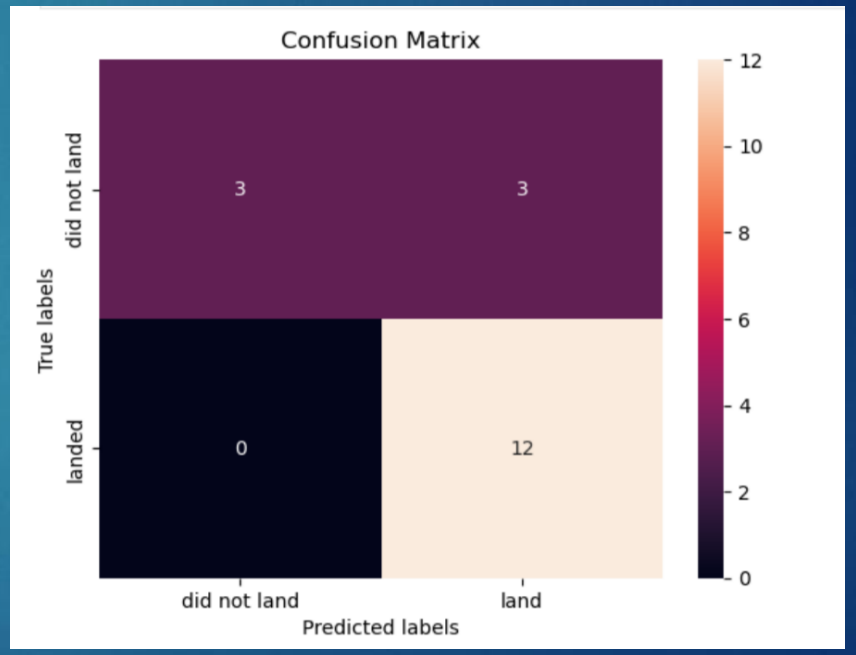
Accuracy

```
accuracy : 0.8464285714285713
```

GridsearchCV score

```
0.8333333333333334
```

Confusion Matrix



Support Vector Machine(SVM)

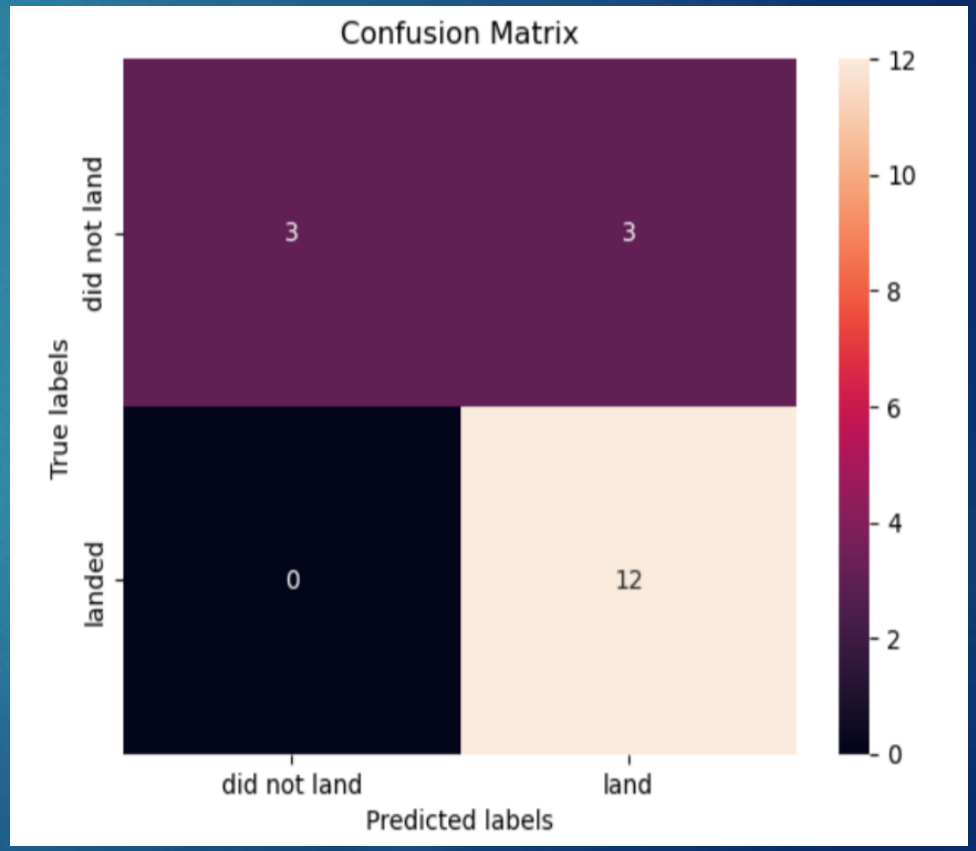
Accuracy

```
accuracy : 0.8482142857142856
```

GridsearchCV Score

```
0.8333333333333334
```

Confusion Matrix



Decision Tree

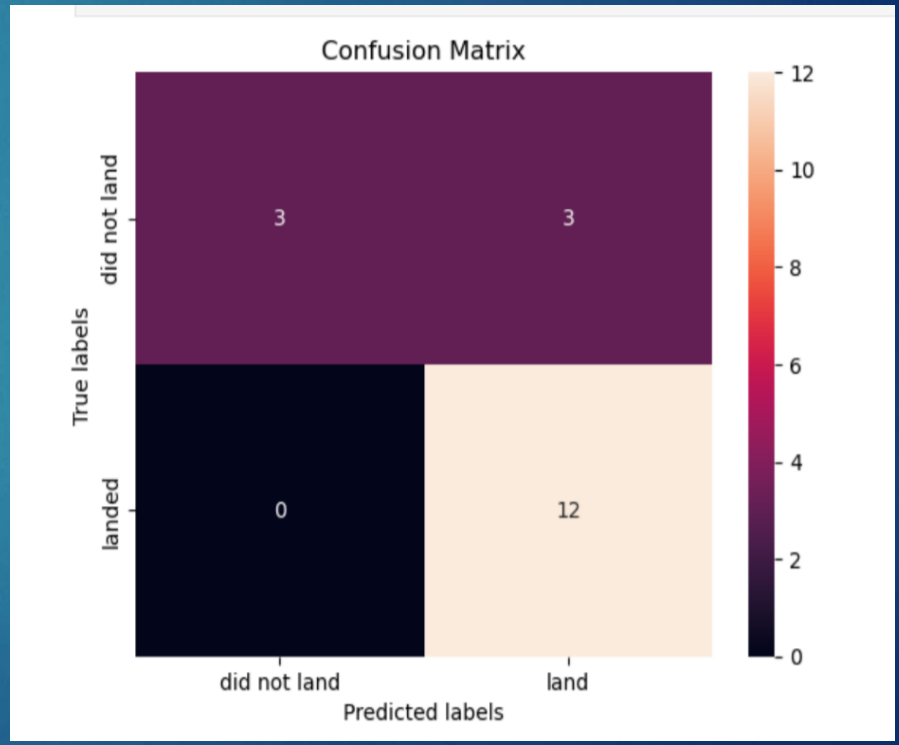
Accuracy

```
accuracy : 0.8857142857142856
```

Gridsearchcv Score

```
0.8333333333333334
```

Confusion Matrix



K-Nearest Neighbors(KNN)

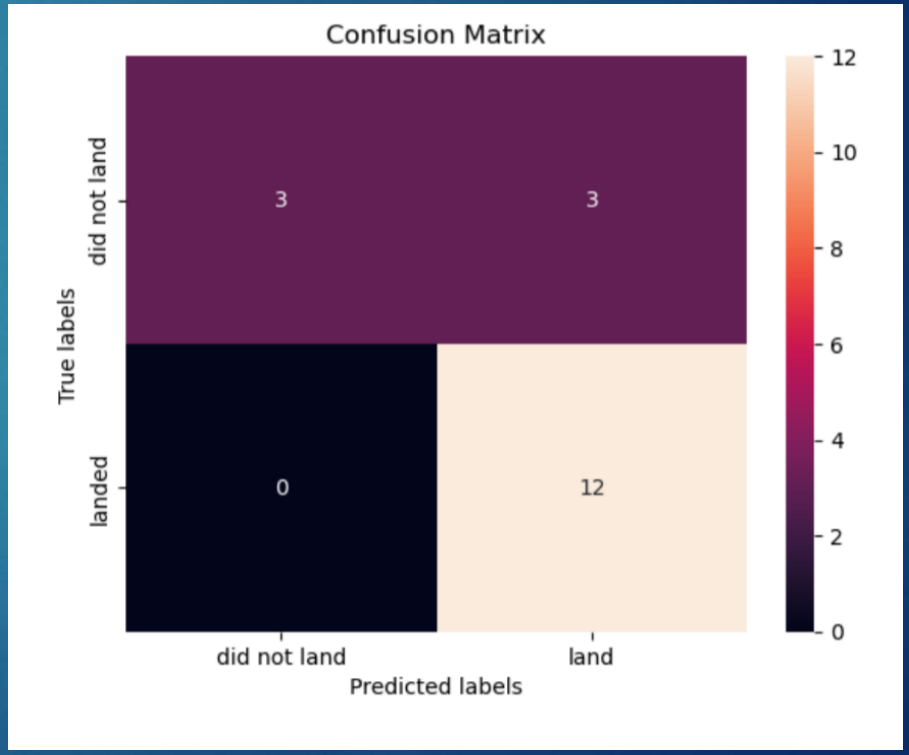
Accuracy

```
accuracy : 0.8482142857142858
```

GridsearchCV Score

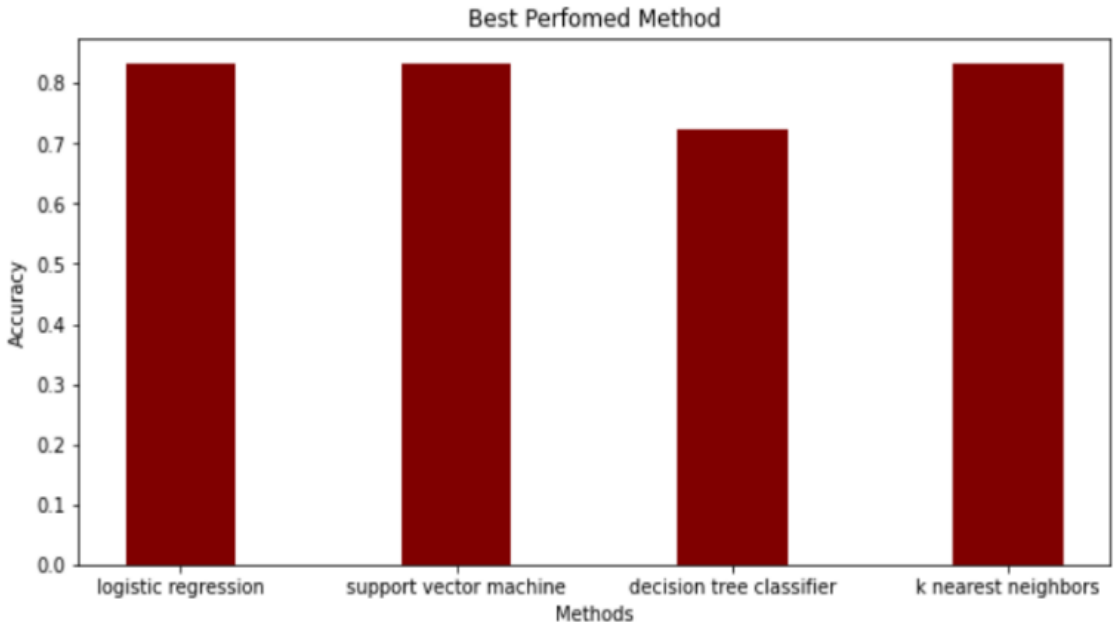
```
0.8333333333333334
```

Confusion Matrix



CLASSIFICATION ACCURACY

Classification Accuracy



DISCUSSION

- From the data visualization, we can see that certain traits may be related to the mission's success in a number of different ways. For instance, with large payloads, the orbit types Polar, LEO, and ISS have higher rates of successful landings or positive landings. However, with GTO, we are unable to make a clear distinction because both positive landing rate and negative landing (mission failure) are present.
- As a result, each characteristic may have a particular effect on the final mission result. It is challenging to determine exactly how each of these features affects the mission's outcome. To learn the pattern of the historical data, however, and forecast the success or failure of a mission based on the provided features, we can use some machine learning techniques.

Contoso

CONCLUSION

This project aims to create a machine learning model for Space Y, which wanted to compete against SpaceX. The model's objective was to predict when Stage 1 would successfully land in order to save $100 million USD. We used data from a public SpaceX API and web-scraped the SpaceX Wikipedia page. We created data labels and stored the data in a DB2SQL database. Finally, we created a dashboard for visualization. In order to decide whether or not to proceed with a launch, Space Y can utilize this model to predict, with a fair amount of accuracy, if a launch will successfully complete a Stage 1 landing before launch. To better choose the optimal machine learning model and increase the accuracy, more data should be gathered.

Contoso

THANK YOU

Sai Durga Vannala

uceku95/ibm-applied-data-science-capstone (github.com)

Contoso