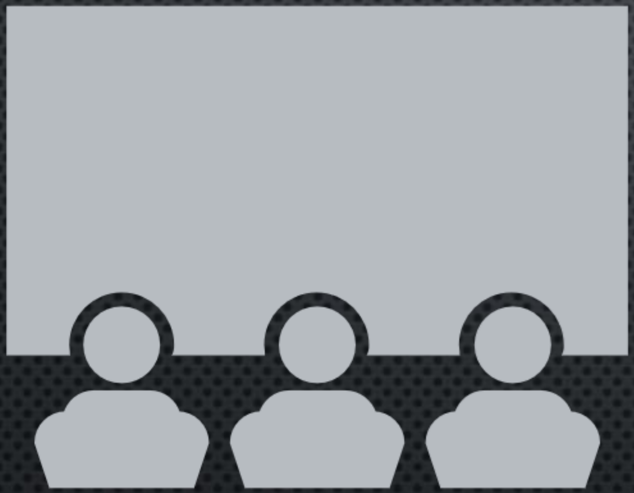
DATA SCIENCE CAPSTONE PROJECT

Vineet Reddy

Date: 18-08-2021

OUTLINE



- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Methodologies: We will first begin with data collection through web scrapping, followed by data processing to clean the data, then we will proceed to data visualization as well as creating our model to predict if the Falcon 9 first stage will land successfully.
- Key results:

(i) Positive relationship between payload mass and success rate of landing

(ii) Most of unsuccessful launches with payload mass under 7000 kg

(iii) KSC LC 39A performs well for payload mass under 5000 kg

(iv) CCAFS SLC 40 performs well for payload mass over 13000 kg

INTRODUCTION

- The commercial space age is here, companies are making space travel affordable for everyone. The most successful is SpaceX.
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; much of the savings is because SpaceX can reuse the first stage.
- If we can determine if the first stage will land, we can determine the cost of a launch.

METHODOLOGY

- Data collection methodology:
  - Used the API to extract information using identification numbers in the launch data.
- Perform data wrangling
  - Dealing with missing values.
- Perform Exploratory Data Analysis (EDA) using visualization and SQL
- Perform Interactive Visual Analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Split into training and test data, tune, evaluate classification models

METHODOLOGY

DATA COLLECTION

- Collected and made sure the data is in the correct format from an SpaceX API.
- Pandas : To collect and manipulate data in JSON and HTML and then data analysis
- requests : Handle http requests
- matplotlib : Detailing the generated maps
- folium : Generating maps of London and Paris
- sklearn : To import Kmeans which is the machine learning model that we are using

DATA COLLECTION – SPACEX API

- https://github.com/Kalyan-A/CapstoneProject/blob/161a9cb75ab8d2bfff6bb01c6f04146a267121ab/Complete%20the%20Data%20Collection%20API%20Lab.ipynb

DATA COLLECTION – WEB SCRAPING

- https://github.com/Kalyan-A/CapstoneProject/blob/161a9cb75ab8d2bfff6bb01c6f04146a267121ab/Complete%20the%20Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

DATA WRANGLING

- There are some missing valuesin the column of ‘PayloadMass’ and ‘LandingPad’, we have to deal with these missing values. For ‘LandingPad’, the missing values will retain none to represent when landing pads were not used. For ‘PayloadMass’, we will replace the missing values with the mean.
- https://github.com/Kalyan-A/CapstoneProject/blob/161a9cb75ab8d2bfff6bb01c6f04146a267121ab/Complete%20the%20Data%20Collection%20API%20Lab.ipynb

EDA WITH DATA VISUALIZATION

Type your text

- https://github.com/Kalyan-A/CapstoneProject/blob/161a9cb75ab8d2bfff6bb01c6f04146a267121ab/Complete%20the%20EDA%20with%20Visualization%20lab.ipynb

EDA WITH SQL

- https://github.com/Kalyan-A/CapstoneProject/blob/161a9cb75ab8d2bfff6bb01c6f04146a267121ab/Complete%20the%20EDA%20with%20SQL%20lab.ipynb

BUILD AN INTERACTIVE MAP WITH FOLIUM

- https://github.com/Kalyan-A/CapstoneProject/blob/161a9cb75ab8d2bfff6bb01c6f04146a267121ab/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

BUILD A DASHBOARD WITH PLOTLY DASH

- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

PREDICTIVE ANALYSIS (CLASSIFICATION)

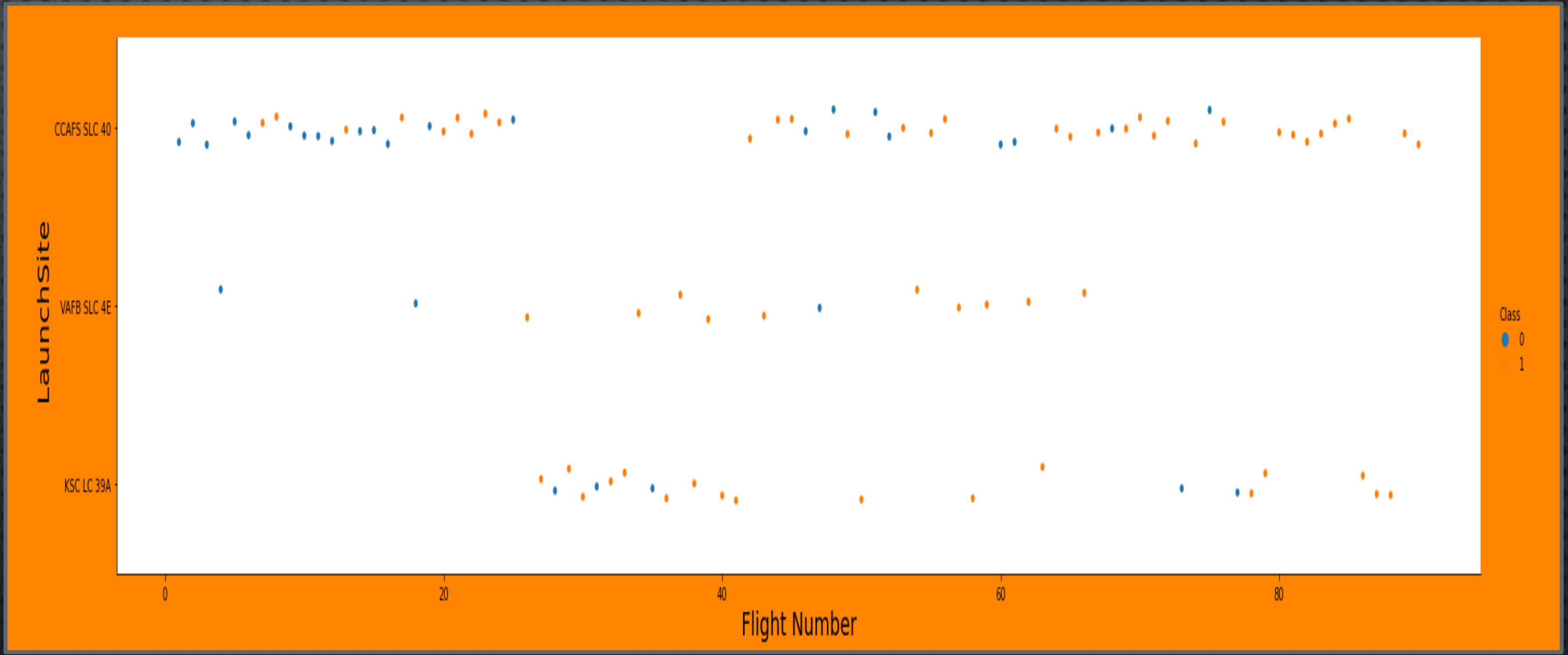
- https://github.com/Kalyan-A/CapstoneProject/blob/161a9cb75ab8d2bfff6bb01c6f04146a267121ab/Complete%20the%20Machine%20Learning%20Prediction%20lab.ipynb

RESULTS

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

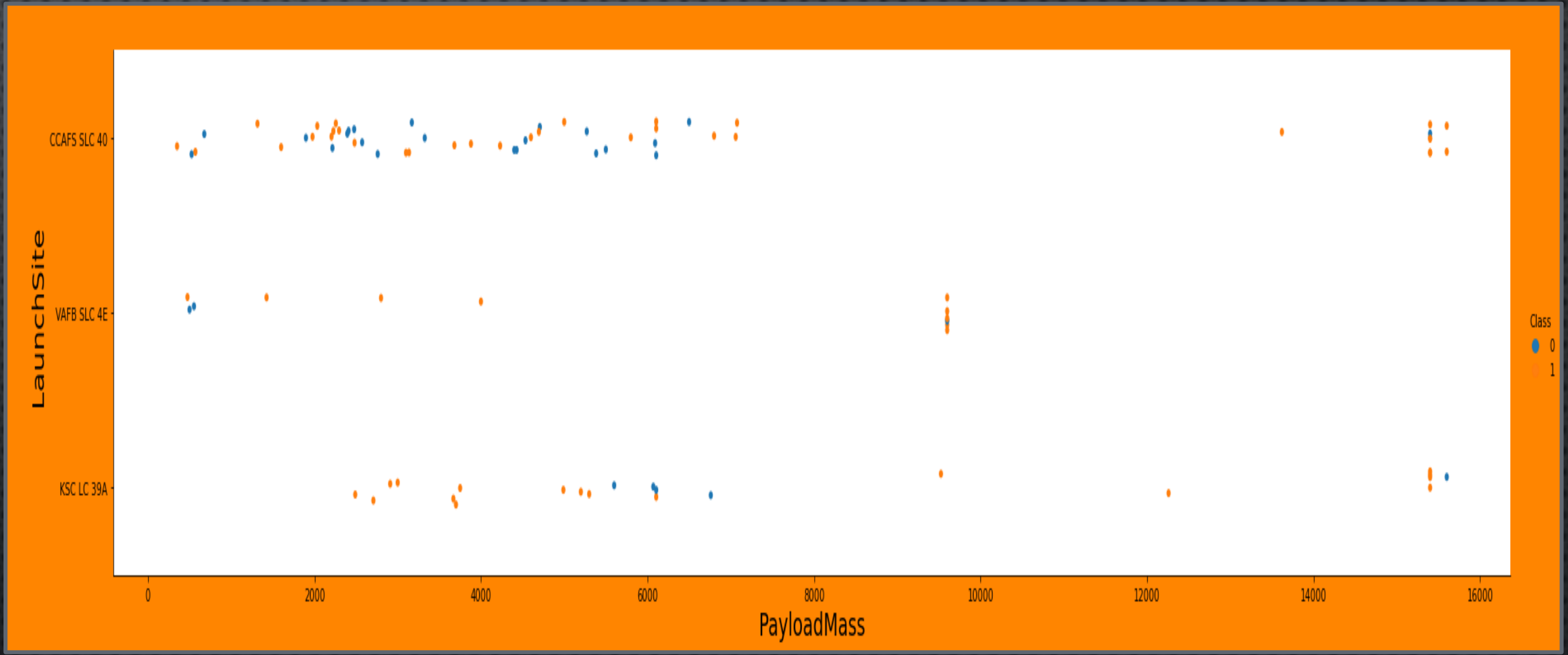
EDA WITH VISUALIZATION

FLIGHT NUMBER VS. LAUNCH SITE



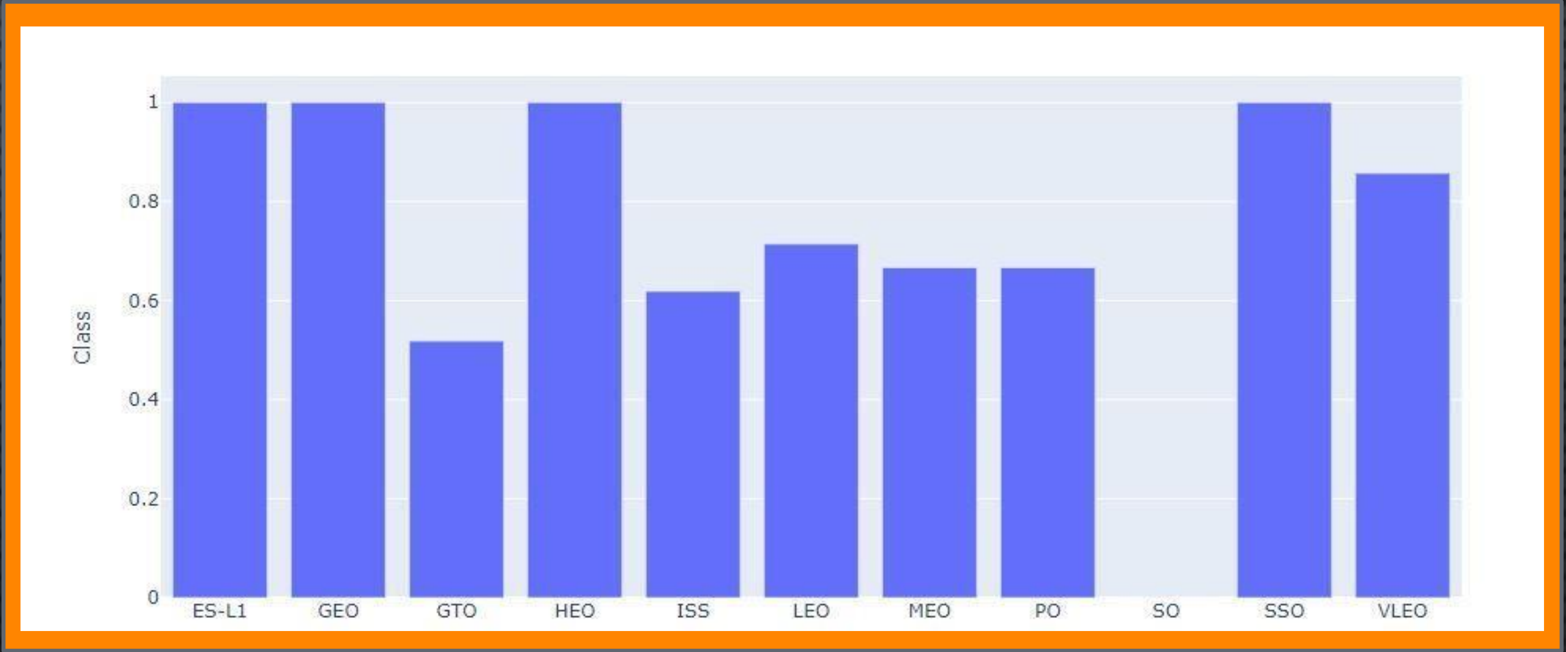
PAYLOAD VS. LAUNCH SITE

scatter plot of Payload vs. Launch Site



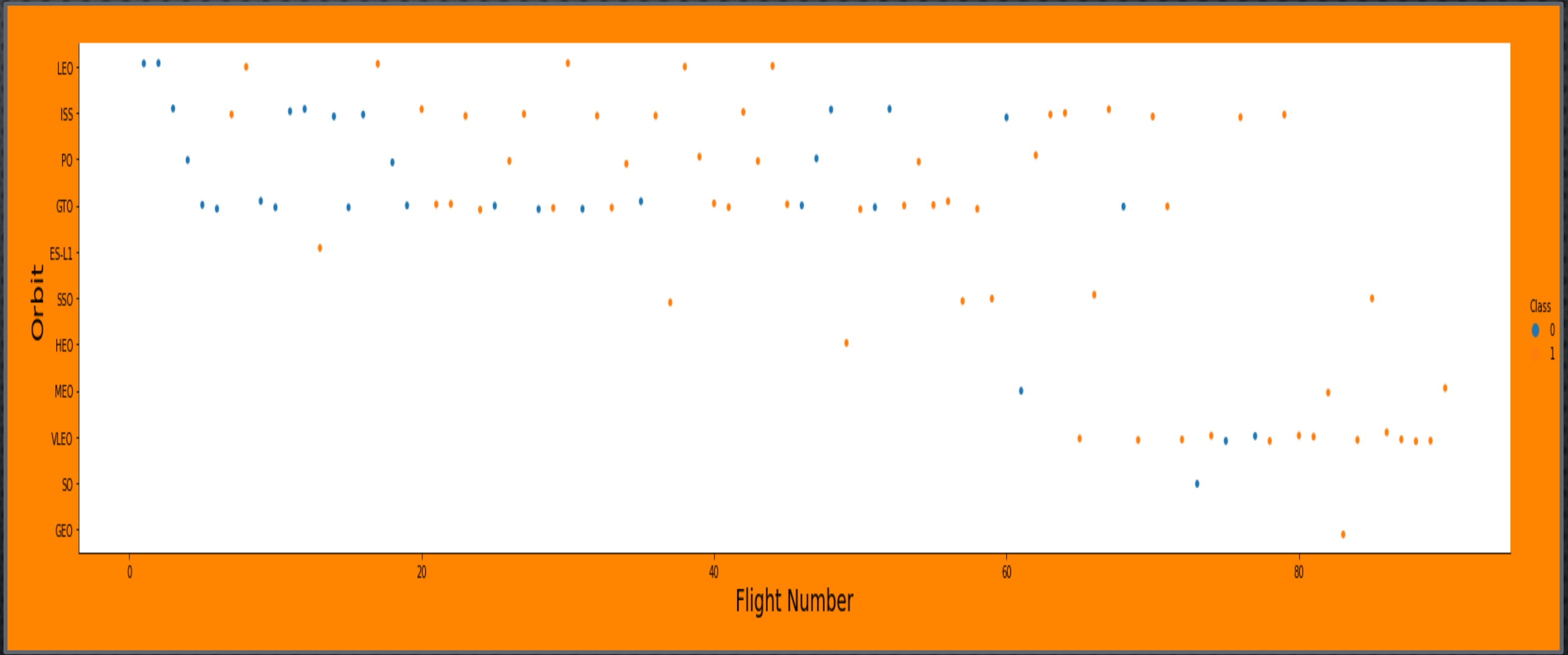
SUCCESS RATE VS. ORBIT TYPE

barchart for the success rate of each orbit type



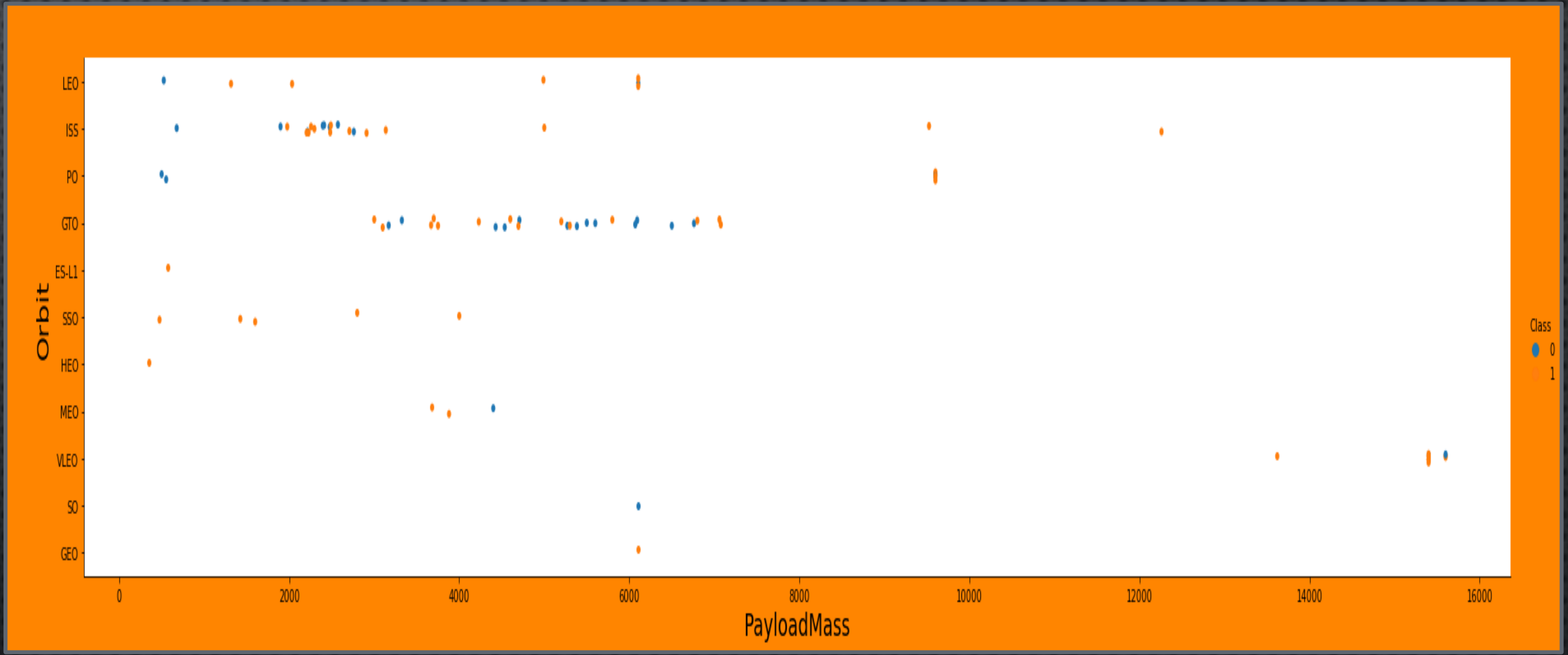
FLIGHT NUMBER VS. ORBIT TYPE

scatter point of Flight number vs. Orbit type



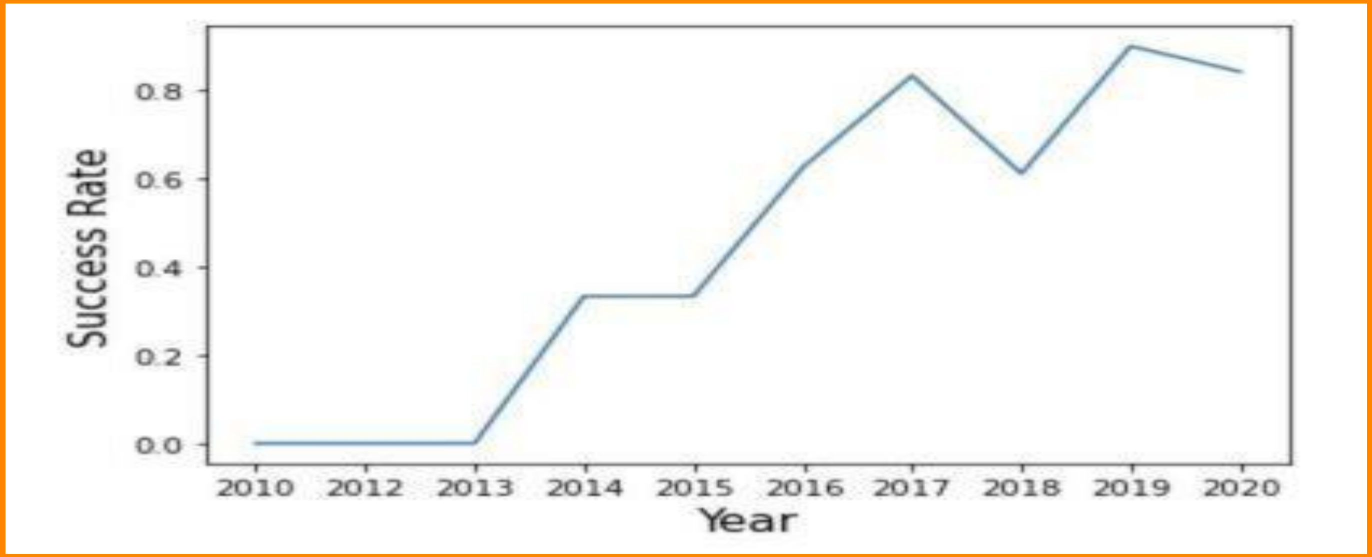
PAYLOAD VS. ORBIT TYPE

scatter point of payload vs. orbit type



LAUNCH SUCCESS YEARLY TREND

line chart of yearly average success rate



EDA WITH SQL

ALL LAUNCH SITE NAMES

- %sql select DISTINCT Launch_Site **from SPACEXTBL**

LAUNCH SITE NAMES BEGIN WITH `CCA`

```
%%sql
SELECT Launch_Site from SPACEXTBL
WHERE Launch_Site  LIKE  'CCA%'
LIMIT 5
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
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)
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt

TOTAL PAYLOAD MASS

- %%sql
  SELECT SUM(PAYLOAD_MASS__KG_) **from SPACEXTBL**
  WHERE Customer='NASA (CRS)';

AVERAGE PAYLOAD MASS BY F9 V1.1

- %%sql

  SELECT AVG(PAYLOAD_MASS__KG_) **from SPACEXTBL**

  WHERE Booster_Version='F9 v1.1';

FIRST SUCCESSFUL GROUND LANDING DATE

- %%sql SELECT MIN(DATE) **from SPACEXTBL**
  WHERE Landing__Outcome='Success (ground pad)';

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD BETWEEN 4000 AND 6000

- %%sql SELECT Booster_Version, PAYLOAD_MASS__KG_ **from SPACEXTBL** WHERE Landing__Outcome= 'Success (drone ship)' AND PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000;

	_

TOTAL NUMBER OF SUCCESSFUL AND FAILURE MISSION OUTCOMES

- %%sql SELECT COUNT(Mission_Outcome) **from SPACEXTBL** WHERE Mission_Outcome LIKE 'Success%'

BOOSTERS CARRIED MAXIMUM PAYLOAD

-

```
%%sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTBL
WHERE Booster_Version= 'F9 v1.1'
```

2015 LAUNCH RECORDS

- %%sql SELECT Month(DATE),Landing__Outcome,Booster_Version,Launch_Site **from SPACEXTBL**

  WHERE Year(DATE)=2015 AND Landing__Outcome= 'Failure (drone ship)' ;

	_		

RANK SUCCESS COUNT BETWEEN 2010-06-04 AND 2017-03-20

EDA with SQL results

Task 10

Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

[22]:

```
%%sql
SELECT landing__outcome, Count(*) AS OUTCOME_COUNT
FROM SPACEXDATASET WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND UPPER(landing__outcome) LIKE 'SUCCESS%'
GROUP BY landing__outcome
ORDER BY OUTCOME_COUNT DESC
```

```
 * ibm_db_sa://xmk21217:***@ba99a9e6-d59e-4883-8fc0-d6a8c9f7a08f.clogj3sd0tgtu0lqde00.databases.appdomain.cloud:31321
/bludb
Done.
```

t[22]:

landing__outcome	outcome_count
Success (drone ship)	5
Success (ground pad)	3

INTERACTIVE MAP WITH FOLIUM

INTERACTIVE MAP WITH FOLIUM

INTERACTIVE MAP WITH FOLIUM

INTERACTIVE MAP WITH FOLIUM

PREDICTIVE ANALYSIS (CLASSIFICATION)

Machine Learning - Logistic Regression

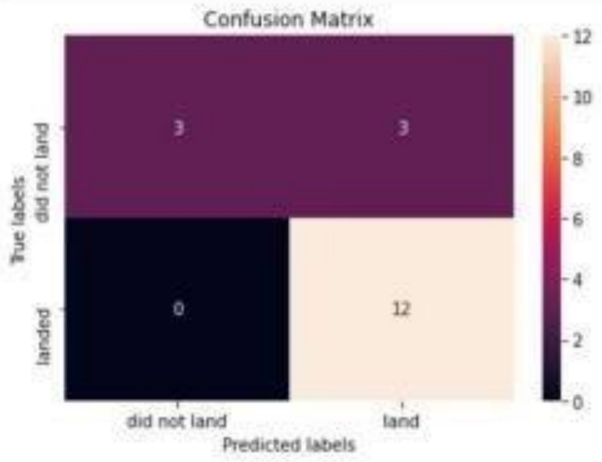
Calculate the accuracy on the test data using the method `score`:

In [14]:
```
logreg_cv.score(X,Y)
```

Out[14]: 0.8666666666666667

Lets look at the confusion matrix:

In [15]:
```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Machine Learning - GridSearch

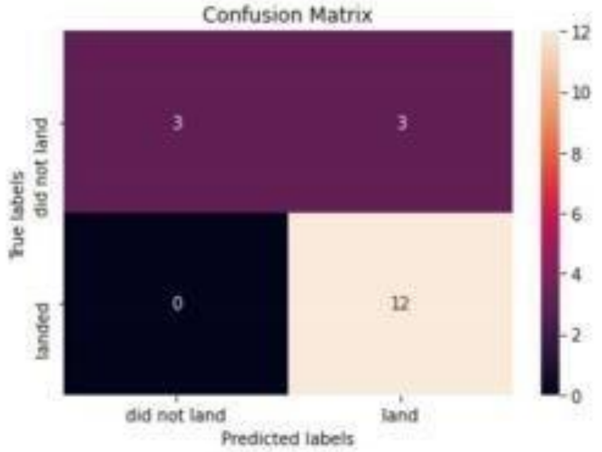
Calculate the accuracy on the test data using the method `score`:

In [19]:
```
svm_cv.score(X,Y)
```

Out[19]: 0.8777777777777778

We can plot the confusion matrix

In [20]:
```
yhat=svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



Machine Learning - Decision Tree

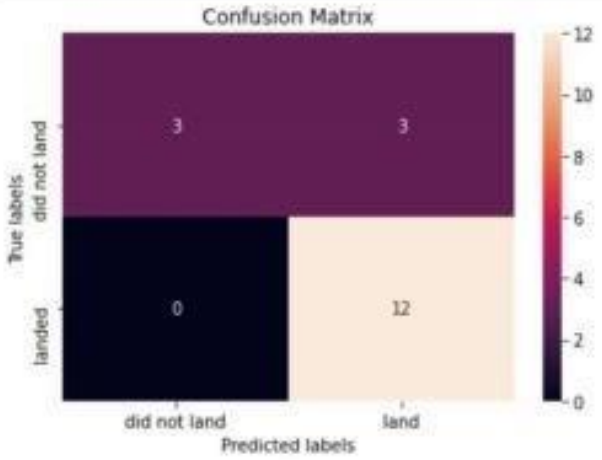
Calculate the accuracy of tree_cv on the test data using the method `score`:

In [24]:
```
tree_cv.score(X,Y)
```

Out[24]: `0.9555555555555556`

We can plot the confusion matrix

In [25]:
```
yhat = svm_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```



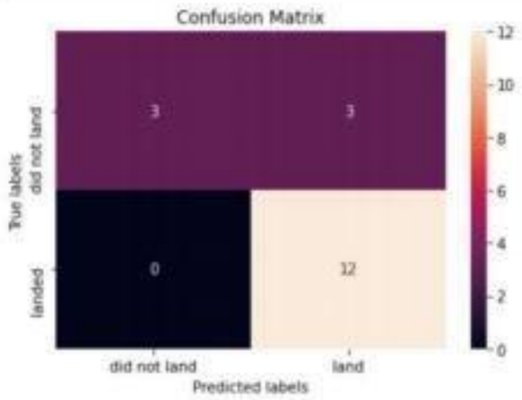
Machine Learning - KNN

Calculate the accuracy of tree_cv on the test data using the method `score`:

```
In [29]: knn_cv.score(X,Y)
Out[29]: 0.8555555555555555
```

We can plot the confusion matrix

```
In [30]: yhat = knn_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



CONCLUSION

- THE PURPOSE OF THE PROJECT IS TO DEVELOP A MODEL WHICH CAN PREDICT THE FIRST STAGE LANDING RESULT AND ALSO IDENTIFY THE RELATIONSHIPS AMONG THE FEATURES.
- BASED ON THE RESULTS, WE FOUND THAT THE DECISON TREE HAS THE
- HIGHEST ACCURACY IN PREDICTING THE LANDING RESULTS.
- APART FROM THIS, WE ALSO IDENTIFY THE FOLLOWING RESULTS BASED ON THE DATA ANALYSIS, (I) POSITIVE RELATIONSHIP BETWEEN PAYLOAD MASS AND SUCCESS RATE OF LANDING, (II) MOST OF UNSUCCESSFUL LAUNCHES WITH PAYLOAD MASS UNDER 7000 KG, (III) KSC LC 39A PERFORMS WELL FOR PAYLOAD MASS UNDER 5000 KG, (IV) CCAFS SLC 40 PERFORMS WELL FOR PAYLOAD MASS OVER 13000