IBM Developer
SKILLS NETWORK

Will the Rocket Fly?

Winning Space Race with Data Science

Dongwei Wu
15 December 2021

Outline

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

Executive Summary

- This project aims to develop an algorithm for predicting the probability of successful rocket launches to help with SpaceY's plan to launch rockets
- Summary of methodologies - I collected data on SpaceX's Falcon 9 rocket launches using SpaceX's API and the Beautiful Soup, processed the data using Numpy, explored the data and extracted features using Matplotlib and Seaborn and finally fed the processed data into various predictive algorithms in the scikit-learn library to determine which algorithm performs the best.
- Summary of all results – the best algorithm as measured by the accuracy score of the algorithm's prediction on test data is the decision tree algorithm with the decision criteria of entropy and maximum depth of 12. The algorithm accuracy on test data is an astonishing 94.44%.

Introduction

- Background – SpaceY is a start-up in the rocket business. It would like to estimate the cost of a launch. A crucial piece of this puzzle is the probability of a successful launch where a part of the rocket gets to be recycled. SpaceY therefore needs a reliable algorithm to predict the probability of such success.
- Problem – an algorithm which can reliably predict the probability of a successful rocket launch whereby the first stage of the rocket is recycled.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data on rocket launches are collected using the request library and SpaceX API from SpaceX's website as well as using the Beautiful Soup library from the following Wikipedia page - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches
- Perform data wrangling
  - A dummy variable "class" is created to signify the success or failure of a launch. Only data relating to launches using a Falcon 9 booster are kept.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data were standardized, split and fed to SVM, decision tree, logistic regression and k-nearest neighbors algorithms through grid-search.

Data Collection

- Data Collection – SpaceX API
- Data Collection - Scraping

Data Collection – SpaceX API

1. Request past launch data using SpaceX API
2. Normalize the response using Panda's Json normalisation method
3. Apply extraction functions to each cell of the data table from and save extracted data into a dictionary
4. Saved the dictionary into a Panda DataFrame

https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project/blob/main/SpaceX%20launch%20data%20collection.ipynb

Data Collection - Scraping

1. Request the raw html file of a Wikipedia page which contains data on past launches of SpaceX
2. Use Beautiful Soup’s findAll method to extract the html relating to all tables
3. Loop through the table html and save the relevant data in a Python dictionary
4. Save the dictionary into a Panda DataFrame

https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project/blob/main/Space%20X%20launch%20data%20collection%20using%20Web-scraping.ipynb

Data Wrangling

- Filter the data to keep only data relating to launches using a Falcon 9 Booster
- Inspect the column “Outcome” and determine which value represents success or failure
- Insert a column named “Class” into the DataFrame which uses 0 and 1 to represent failure and success respectively

https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project/blob/main/EDA%20(Data%20Wrangling).ipynb

EDA with Data Visualization

- Scatter Plot of Flight Number vs. Payload Mass
- Scatter Plot of Flight Number vs. Launch Sites
- Scatter Plot of Payload Mass vs. Launch Site
- Bar Chart showing success rate of each orbit
- Scatter Plot of Flight Number vs. Orbit
- Scatter Plot of Payload Mass vs. Orbit
- Line Chart of Year vs. Success Rate

https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project/blob/main/EDA%20with%20Data%20Visualization.ipynb

EDA with SQL

- Display distinct launch sites
- Display only launch sites in Cape Canaveral Space Force Station, Florida
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster versions which have carried the maximum payload mass (sub-query used)
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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

https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project/blob/main/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Add circles to the locations of each launch sites to indicate location as well as markers to indicate the names of the launch sites
- Add marker clusters to indicate the number of launches, successes and failures
- Add MousePosition to show the coordinate of where the mouse is on a map to facilitate the calculation of the proximity between the launch site and various landmarks
- Add polyline to mark the straight line distance between certain launch site and its nearest coast, highway and city center

https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project/blob/main/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Pie Chart to show the split of launches across all launch sites
- Pie Chart to show the rate of success/failure in each launch site
- Scatter Plot to show the correlation between success and payload mass with different color for each booster version to show its correlation with success
- Summarize what plots/graphs and interactions you have added to a dashboard
- https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project/blob/main/Interactive%20Dashboard.ipynb

Predictive Analysis (Classification)

1. Load the DataFrame and split it into the feature set and the label set
2. Standardize the feature set using sklearn’s StandardScaler
3. Split the dataset into the training set and testing set with a 80:20 ratio
4. Feed the data into four learning algorithms – SVM, K-nearest neighbours, logistic regression and decision tree.
5. For each algorithm, a grid search object was created so that different parameters are explored
6. The best parameters for each algorithm are found and the model was tested against the testing set
7. The model which performs the best on the testing set in terms of predictive accuracy is selected

https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project/blob/main/Predictive%20Analysis%20(Classification).ipynb

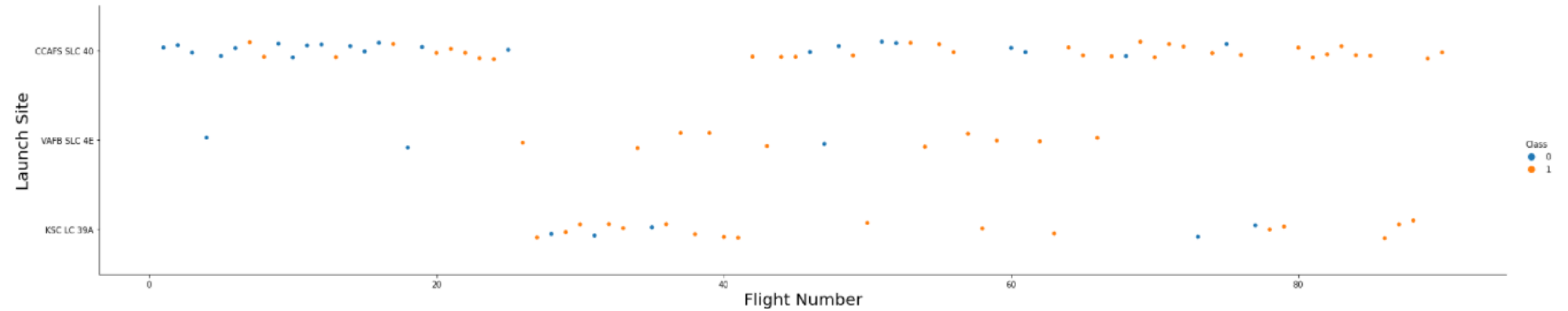
Results

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

Section 2

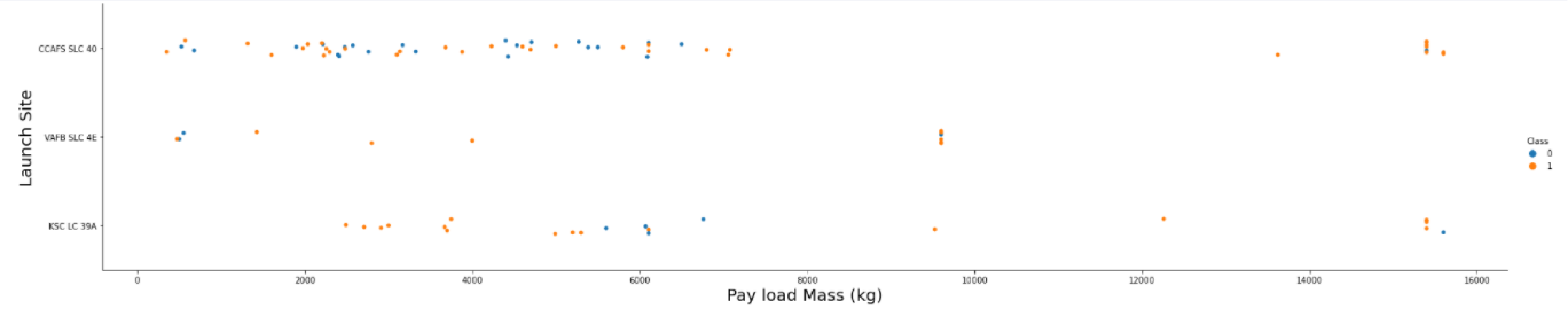
Insights drawn from EDA

Flight Number vs. Launch Site



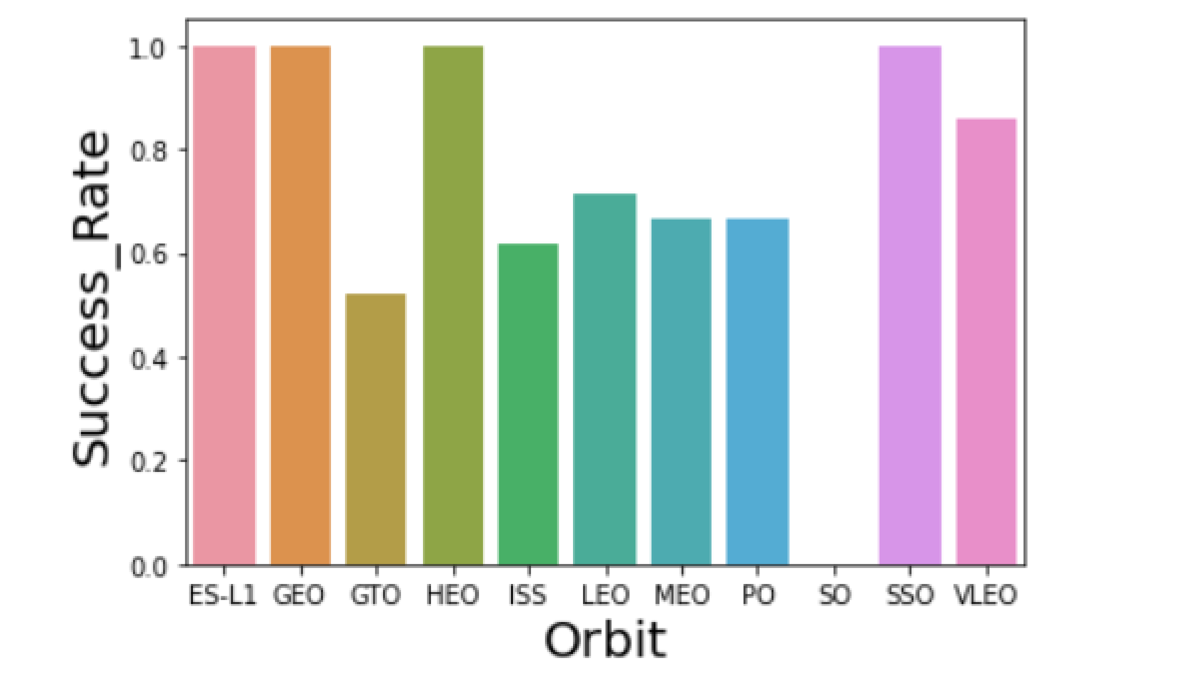
- CCAFS SLC 40 has the greatest number of launches
- Launches 1 to 25 and 40 to over 80 tend to happen in CCAFS SLC 40
- Launches 25 to 40 tend to happen in KSC LC 39A

Payload vs. Launch Site



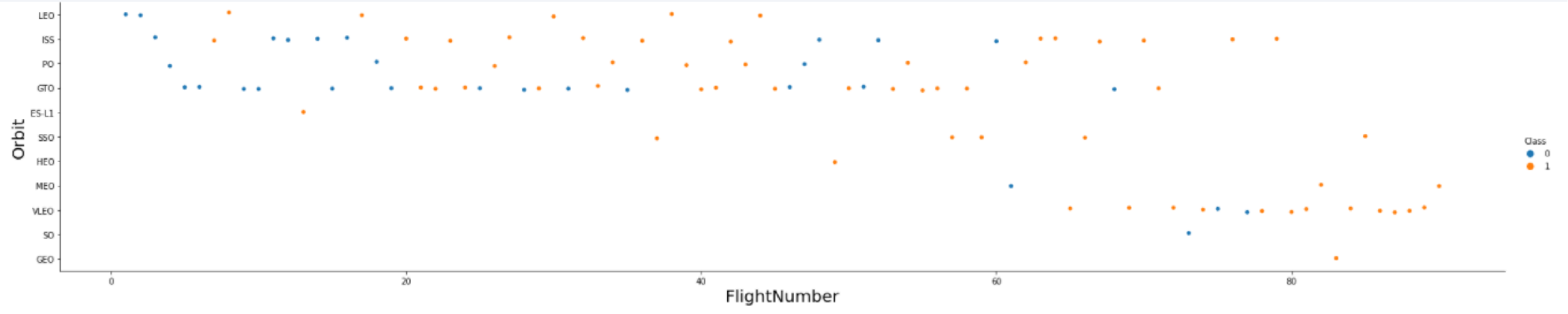
- The heaviest pay loads are launched across CCAFS SLC 40 and KSC LC 39A, but not VAFB SLC 4E, which tends to launch payloads which are slightly below 10000kg
- CCAFS SLC 40 handles a range of payloads from a few hundred kg to around 7000kg
- At the lower end, KSC LC 39A tends to handle payloads between 2000kg and 7000kg

Success Rate vs. Orbit Type



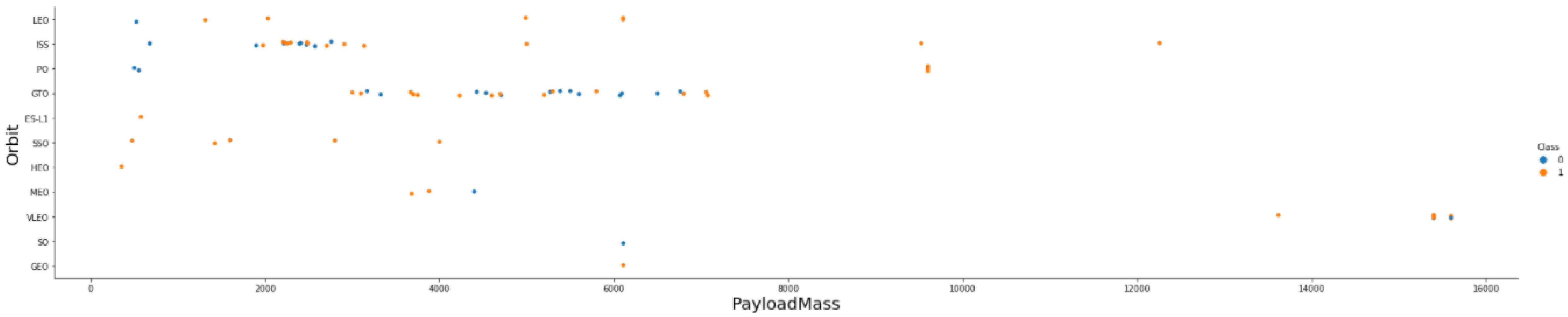
- Success rates for ES-L1, GEO, HEO and SSO are the perfect 100%
- Success rate for GTO is the lowest, a 50% coin-flip,
- Success rates for ISS, MEO, PO and LEO are between 60% and 70%

Flight Number vs. Orbit Type



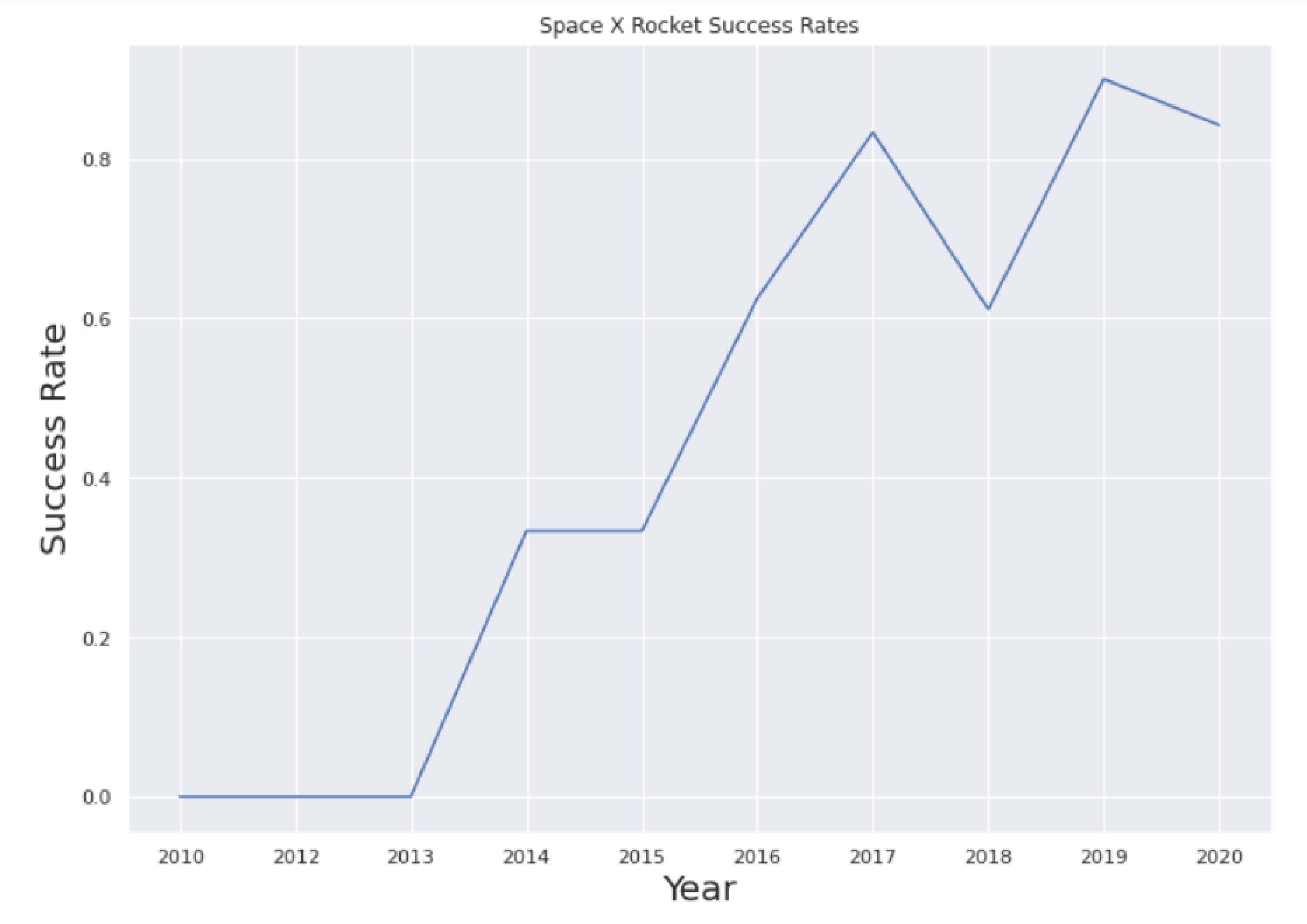
- The most frequently attempted orbit is GTO
- Certain orbits are only attempted after 60 flights, such as VLEO, SO, GEO and MEO

Payload vs. Orbit Type



- The heaviest payloads (over 13,000kg) are all carried out on VLEO
- The widest range of payloads are carried out on ISS (from 500kg to over 12,000kg)

Launch Success Yearly Trend



- The success rate increases in a linear fashion year-on-year from all failures to over 85% successes

All Launch Site Names

Task 1

Display the names of the unique launch sites in the space mission

In [5]:
```
%sql select distinct(LAUNCH_SITE) from SPACEXDATASET
```

```
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[5]:

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

- '%sql' allow jupyter notebook to process SQL query
- 'Select' and 'from' returns data from a chosen dataset
- 'distinct(column_name)' keeps only the unique values in the column

Launch Site Names Begin with 'CCA'

In [7]:
```
%sql select * from SPACEXDATASET where LAUNCH_SITE like 'CCA%' limit 5
```

```
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[7]:

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

- ‘where’ specifies conditions for filtering the selected dataset
- Column_name ‘like’ ‘value’ retains entries whose values follow the format in the specified value. ‘string%’ represents any value which start with ‘string’
- ‘limit’ keep the displayed data to only the top five rows selected

Total Payload Mass

```
In [8]:  %sql select sum(PAYLOAD_MASS__KG_) from SPACEXDATASET where CUSTOMER = 'NASA (CRS)'

 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.

Out[8]:    1
        45596
```

- ‘sum(column_name)’ return the sum of all values in the column
- ‘= string’ evaluate whether the value matches exactly with ‘string’, in this case, the query only concerns data where the customer is NASA (CRS)

Average Payload Mass by F9 v1.1

```
In [12]: %sql select avg(PAYLOAD_MASS__KG_) from SPACEXDATASET where booster_version like 'F9 v1.1%'

 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.

Out[12]:    1
         2534
```

- ‘avg(column_name)’ return the average of all values in the column
- This query returns the average payload of all launches powered by the F9 v1.1 boosters

First Successful Ground Landing Date

```
%sql select min(DATE) from SPACEXDATASET where LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

1
2015-12-22

- 'min(column_name) returns the smallest value in the column
- This query returns the earliest launch with a successful ground landing

Successful Drone Ship Landing with Payload between 4000 and 6000

```
]: %sql select BOOSTER_VERSION from SPACEXDATASET where Landing__Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000
```

```
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

19]:

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

- This query lists the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000
- ‘and’ adds further condition to the selection of data

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(MISSION_OUTCOME) from SPACEXDATASET where MISSION_OUTCOME like 'Success%'
```

```
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

1
100

```
%sql select count(MISSION_OUTCOME) from SPACEXDATASET where MISSION_OUTCOME like 'Failure%'
```

```
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

7]:

1
1

- These queries respectively return the total number of launches which succeeded and failed in completing the mission

Boosters Carried Maximum Payload

```
In [30]: %sql select BOOSTER_VERSION from SPACEXDATASET where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXDATASET)

 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.

Out[30]:
```

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

- List of boosters which have carried the maximum payload in all launches
- ‘Max(column_name)’ returns the highest value in the column

2015 Launch Records

```
%sql select LANDING__OUTCOME,BOOSTER_VERSION, LAUNCH_SITE from SPACEXDATASET where LANDING__OUTCOME = 'Failure (drone ship)' and DATE like '2015%'
```

```
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

]:

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

In [42]:

```
%sql select * from SPACEXDATASET where LANDING__OUTCOME like 'Success%' and (DATE between '2010-06-04' and '2017-03-20') order by DATE desc
```

```
 * ibm_db_sa://sfy60079:***@55fbc997-9266-4331-afd3-888b05e734c0.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31929/bludb
Done.
```

Out[42]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2017-01-14	17:54:00	F9 FT B1029.1	VAFB SLC-4E	Iridium NEXT 1	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
2016-08-14	05:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-07-18	04:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-05-06	05:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-04-08	20:43:00	F9 FT B1021.1	CCAFS LC-40	SpaceX CRS-8	3136	LEO (ISS)	NASA (CRS)	Success	Success (drone ship)
2015-12-22	01:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)

- 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

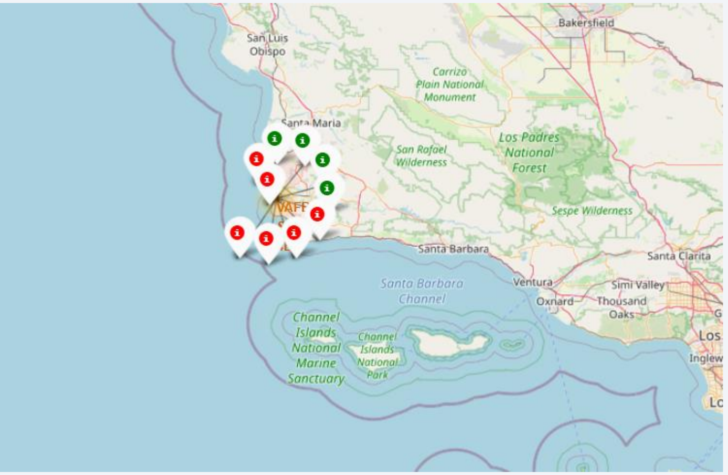
Section 4

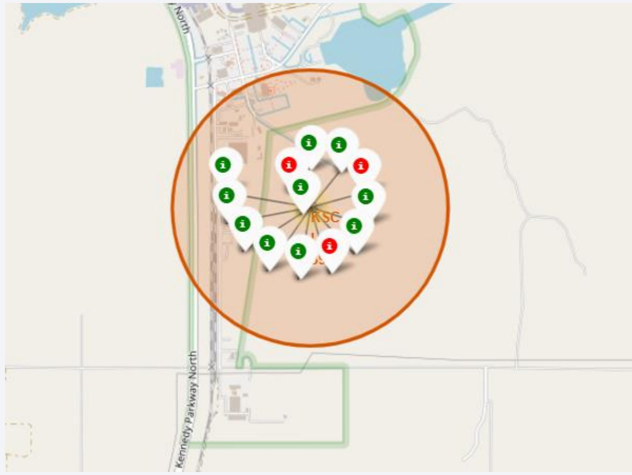
Launch Sites Proximities Analysis

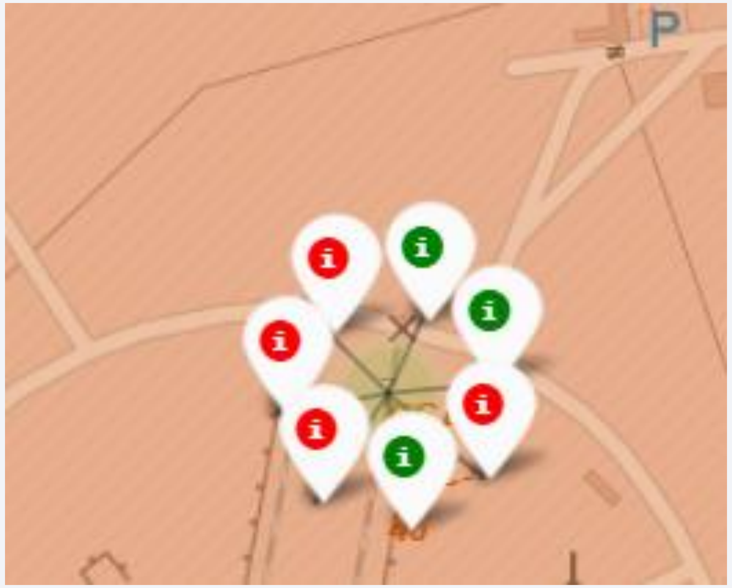
Locations of Launch Sites

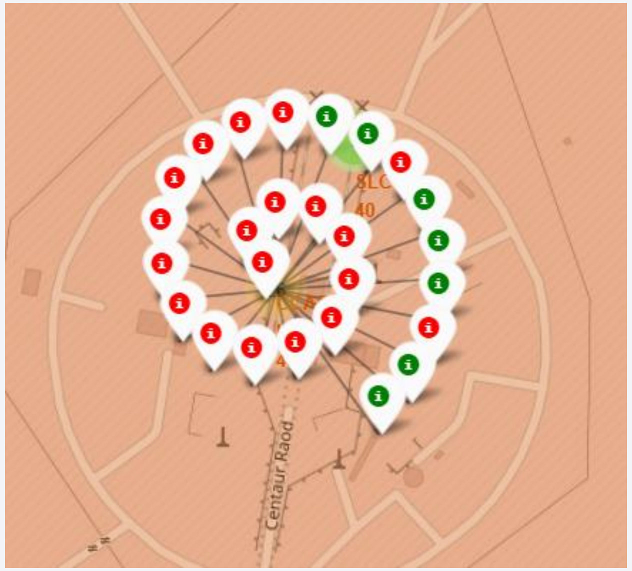
- All launch sites are set up in the U.S.A. and near a coast

Launch Successes and Failures by Site



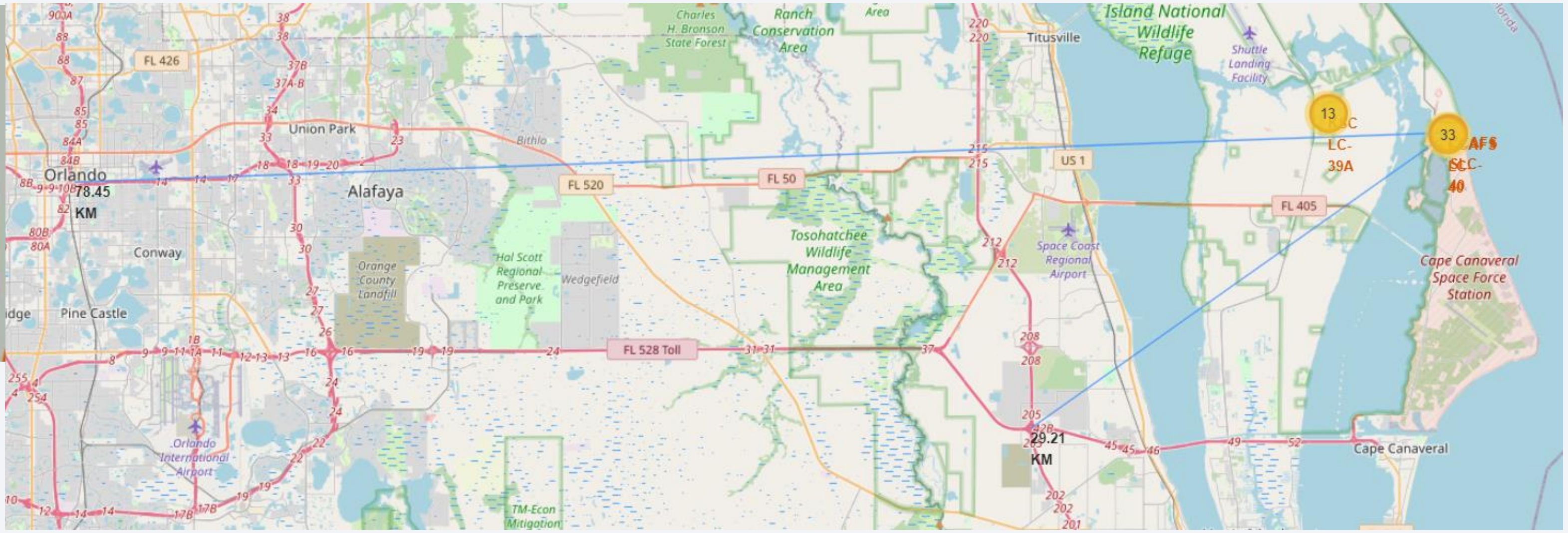






- CCAFS SLC-40 has the greatest number of launches but more failure than success
- KSC LC-39A has the greatest success rate

Proximity to coastlines, highways and cities

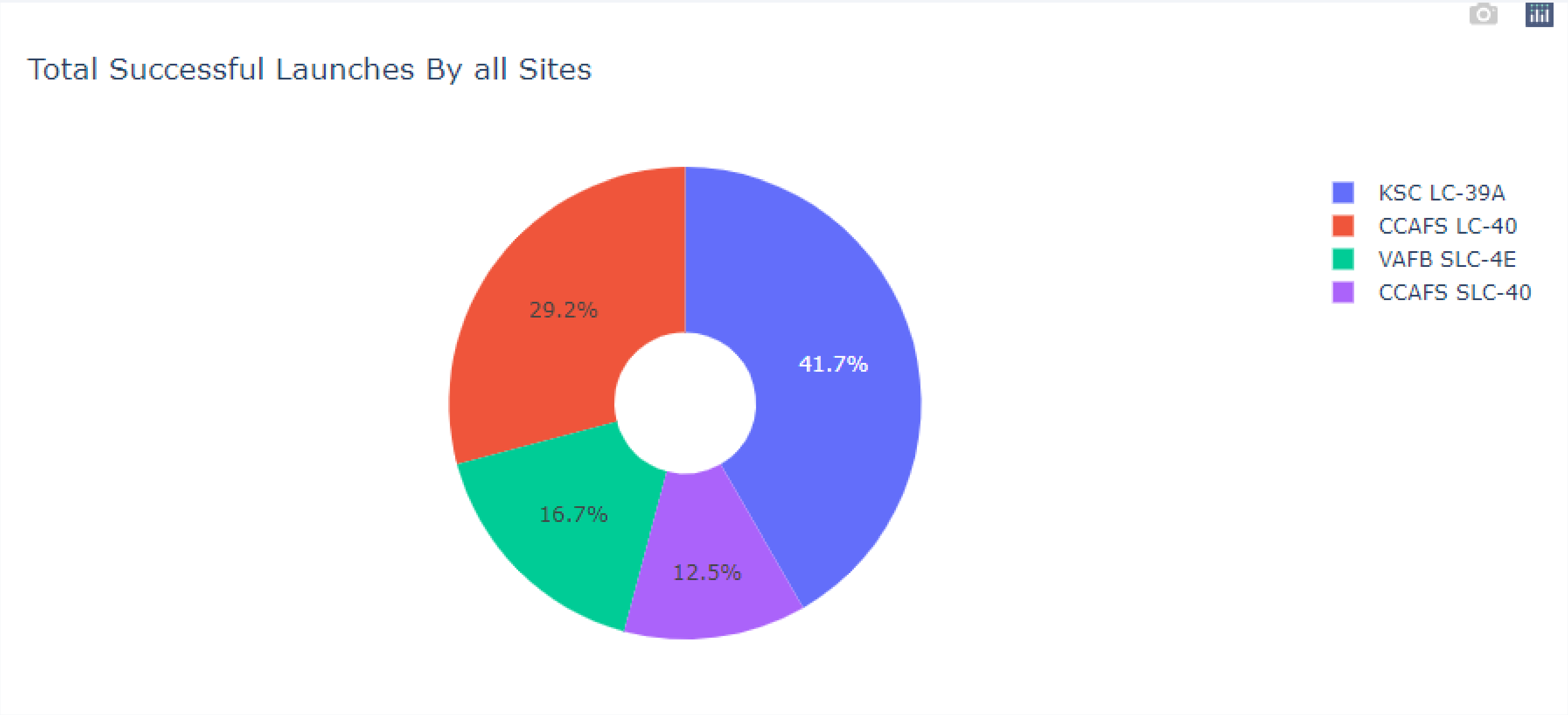


- Launch sites are built close to the coastline and far away from the highway and city center

Section 5

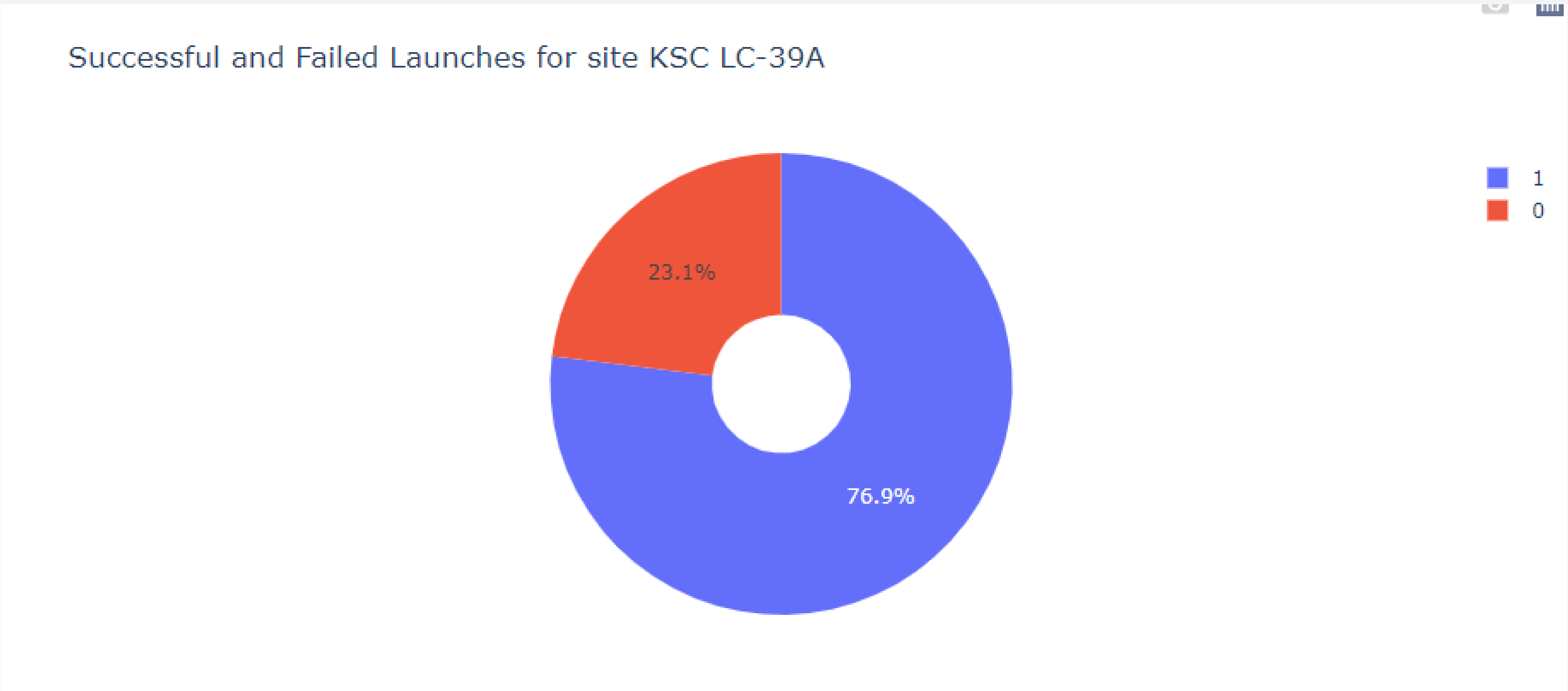
Build a Dashboard with Plotly Dash

Total Successful Launches by Site



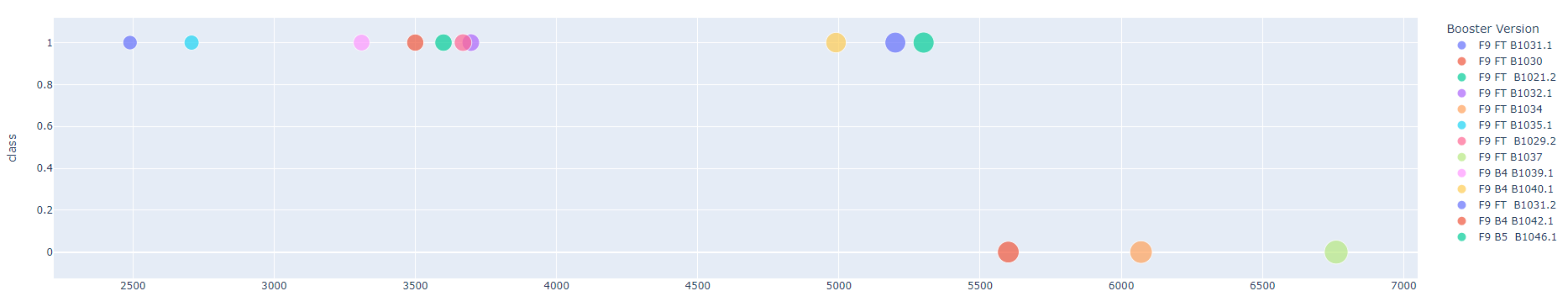
- KSC LC-39A hosted the greatest proportion of successful launches at 41.7% whereas CCAFS SLC-40 hosted the smallest proportion at 12.5%

Most Successful Launch Site by Success Rate



- Close to 77% of the launches at KSC LC-39A are successful

Success Rates of Different Boosters for different payloads

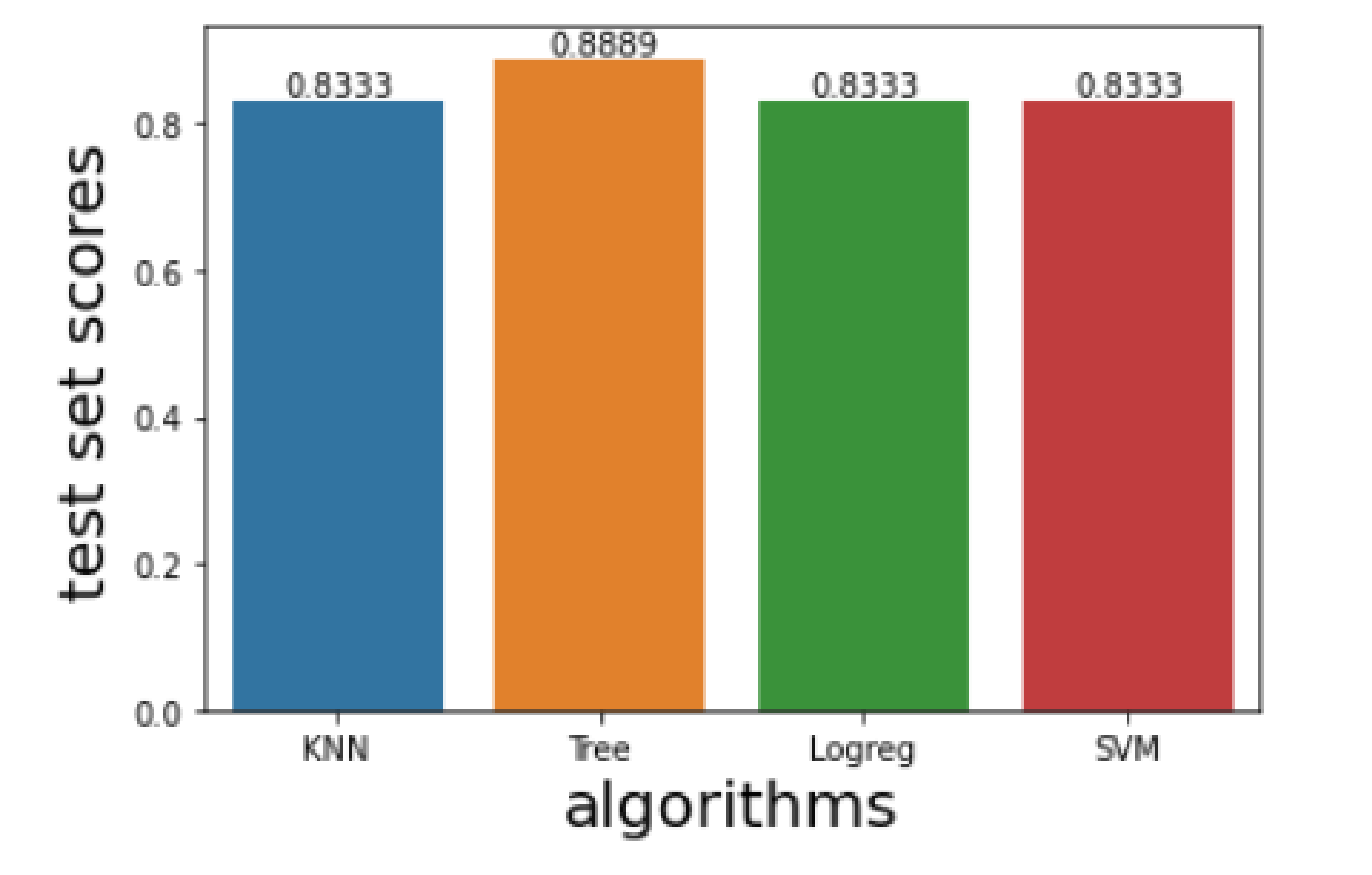


- Each payload weight requires a different sub-version of booster.
- The F9 FT booster covers the widest range of payload mass, from 2500kg to 6750kg
- The F9 B5 booster has only been used for one payload mass at around 5300kg

Section 6

Predictive Analysis (Classification)

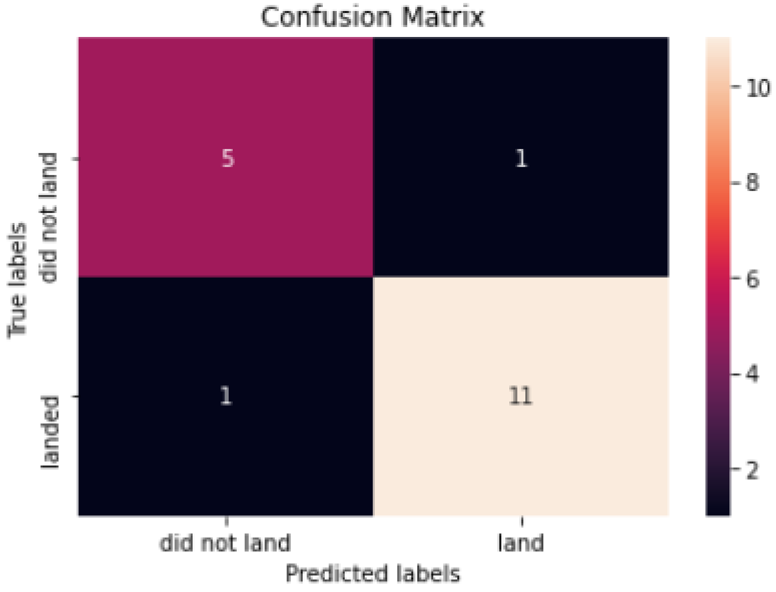
Classification Accuracy



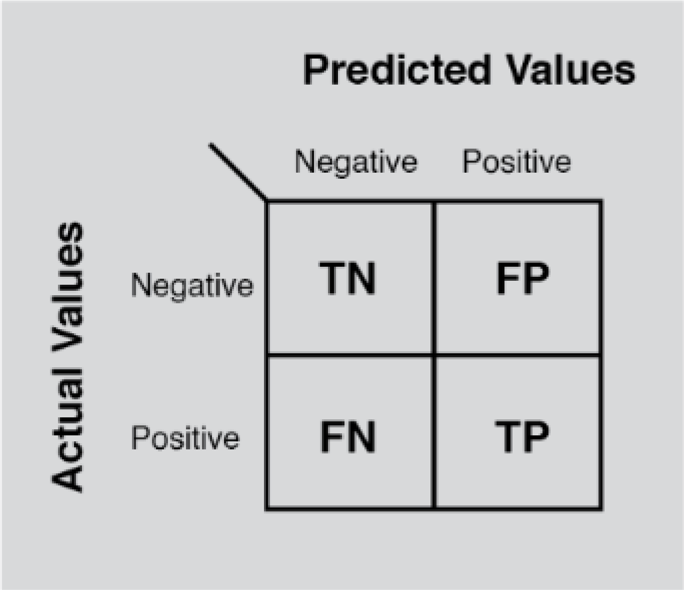
- The decision tree model has the highest accuracy rate of approximately 88.89%

Confusion Matrix of the Decision Tree Model

```
In [43]: yhat = tree_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



- The false positive rate is 16.66% being the higher classification error rate compared with the false negative rate of 8.33%



Conclusions

- The decision tree model is the best for predicting the probability of successful landing of the first stage of a rocket launch – its main weakest is spotting false positive, i.e. predicting a success when the launch was a failure. However, the algorithms' predictive accuracy is only marginally lower than the decision tree model's
- The launch site which contributed the most successful launches and have the highest successful launch rate is KSC LC-39A
- Probability of success increases as payload mass decreases and as flight number increases
- Launches aiming at these orbits are the most successful: ES-L1, GEO, HEO and SSO

Appendix

- All codes, datasets can be found on my Github at https://github.com/azurecode1119/IBD-data-science-professional-certificate-capstone-project

Thank you!