IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mehmet Fettahoglu
12/8/2022

Outline

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

Executive Summary

- In this project, machine learning models were developed to predict the success of landing of the first stage of a rocket launch using data from Space X.
- 4 different machine learning models were trained for modeling: Logistic Regression, SVM, Decision Tree, KNN.
- The results have indicated that all four models yield the same prediction accuracy.

Introduction

- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch.

- Using parameters such as payload, launch site, orbit and others we would like to determine if the first stage of a launch will land back successfully, and use this information to determine the price of a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using the GET request method data was collected from SpaceX
- Perform data wrangling
  - Missing values were filled with column mean, non useful columns were removed, binary variables were created
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Logistic Regression, SVM, Decision Tree, KNN

Data Collection

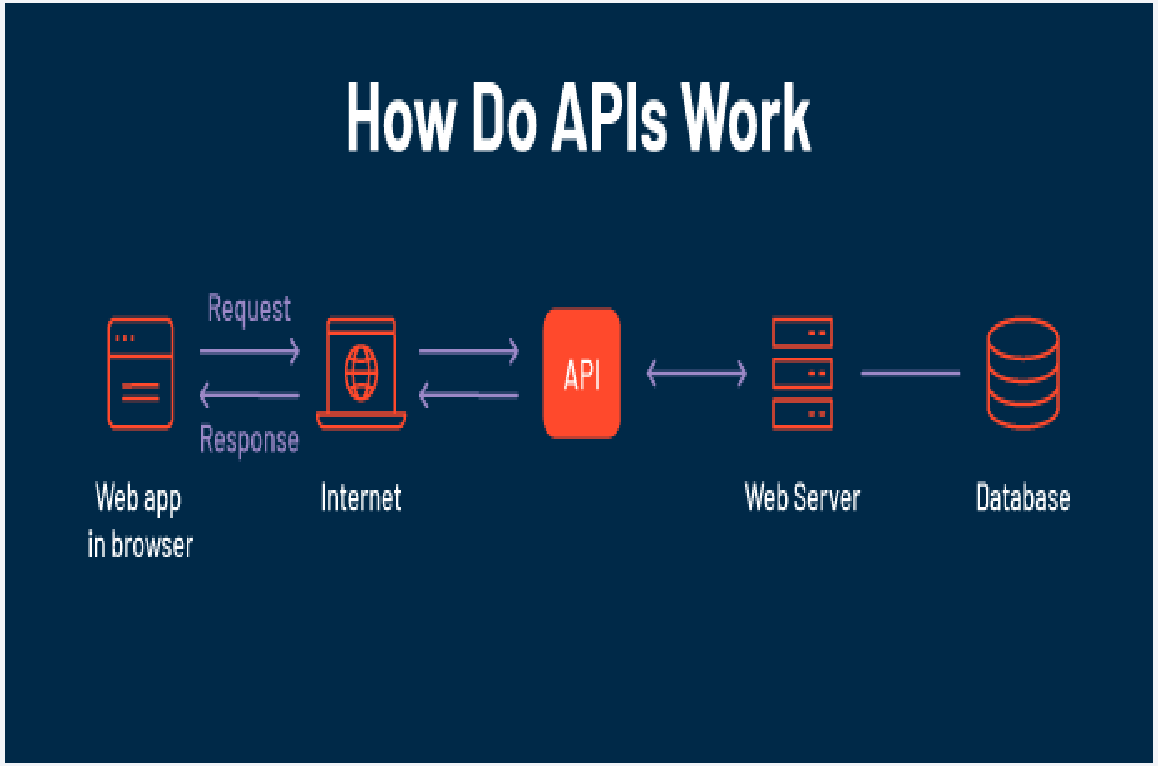
- Space X rocket launch data was collected with a GET request. The response was decoded as a JSON file and turned into a pandas dataframe

- Falcon 9 launch records HTML table from Wikipedia was extracted with GET request, a BeautifulSoup object was created from response, and the table was parsed and turned into a dataframe

Data Collection – SpaceX API

- Rocket launch data was collected from the SpaceX API using requests.get(), and it was decoded using .json_normalize()

- https://github.com/mehmetbe/IMB-DS-Capstone/blob/main/1-jupyter-labs-spacex-data-collection-api.ipynb



Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from its URL - Extract all column/variable names from the HTML table header - Create a data frame by parsing the launch HTML tables

- https://github.com/mehmetbe/IMB-DS-Capstone/blob/main/2-jupyter-labs-webscraping.ipynb

Data Wrangling

- Data was filtered to only include Falcon 9 launches.
- Missing values were replaced by the column mean.
- The number and occurrence of mission outcome per orbit type was calculated, landing outcome variable was created where successful outcomes take the value of 1 and 0 otherwise.
- Dummy variables were created for categorical variables such as orbits, launch site, landing pad, and serial
- https://github.com/mehmetbe/IMB-DS-Capstone/blob/main/3-labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Flight Number (indicating continuous launch attempts) was plotted against Payload Mass with Outcome as color
- Flight Number was plotted against Launch Site with Outcome as color
- Payload mass was plotted against Launch Site with Outcome as color
- Success rate of each orbit was plotted in a bar chart
- Flight Number was plotted against Orbit Type with Outcome as color
- Payload was plotted against Orbit Type with Outcome as color
- Launch success yearly trend was plotted in a line chart
- https://github.com/mehmetbe/IMB-DS-Capstone/blob/main/5-jupyter-labs-eda-dataviz.ipynb

EDA with SQL

SQL queries were created to:

- Display names of unique launch sites
- Display 4 records where launch sites begin with 'CCA'
- 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 achieved
- 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. Use a subquery
- List the records which will display the month names, failure landing outcomes in drone ship, booster versions, and launch site for the months in year 2015.
- Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.
- https://github.com/mehmetbe/IMB-DS-Capstone/blob/main/4-jupyter-labs-eda-sql-coursera_sqllite.ipynb

Build an Interactive Map with Folium

- All launch sites were marked using markers and circles, the success/fail launches were added on each site using markers and clustered using cluster objects, the distances between a launch site and its proximities (e.g. cities or highways) were added using markers and polyline objects, and a mouseposition object was added to get coordinates for a mouse over a point.
- These objects were added in order to visualize certain spots on the map.
- https://github.com/mehmetbe/IMB-DS-Capstone/blob/main/6-lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Dropdown list added to enable launch site selection
- Pie chart was added to show the total successful launches count
- Slider added to select payload range, scatter chart added to show correlation between payload and launch success
- Callback functions added for site selection as input and pie chart as output & site selection and payload slider as input and success scatterplot as output
- https://github.com/mehmetbe/IMB-DS-Capstone/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

- Data was standardized and split to training and test sets (20% test size)
- Logistic regression model was developed wit cv=10, penalty=l2, solver= lbfgs
- SVM model was created with cv =10
- Decision tree model was created with cv =10
- K nearest neighbots model was created with cv=10
- Confusion matrices and accuracy scores were created for each model
- All models gave the same confusion matrices and same accuracy score of 83%
- https://github.com/mehmetbe/IMB-DS-Capstone/blob/main/7-SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

EDA results:

- Launch success increases as flight number increases, indicating with more flights the more experienced Space X got and success rate got better.
- Most unsuccessful outcomes occurred between payload mass of 4000-8000 kg
- VAFB SLC 4E didn’t have any rockets launching greater than 10000 kg payload mass
- Success rate of 4 orbits (ES-L1, HEO, GEO, SSO) was 100%, making them highest successful. Lowest success rate orbits were SO (0%), and GTO (50%)
- Certain orbits haven’t been used past a certain flight no (e.g. LEO after flight 50). LEO orbit success correlated with flight number, which wasn’t the case in VLEO. Orbits like ISS and GTO have been used for a long time period.
- With heavy payloads the successful landing or positive landing rate are higher for Polar, LEO and ISS.
- Success rate since 2013 kept increasing till 2020

Predictive analysis results:

- Confusion matrix for all four models showed all models found 12 true positives, 3 false positives, 0 false negatives, and 3 true negatives

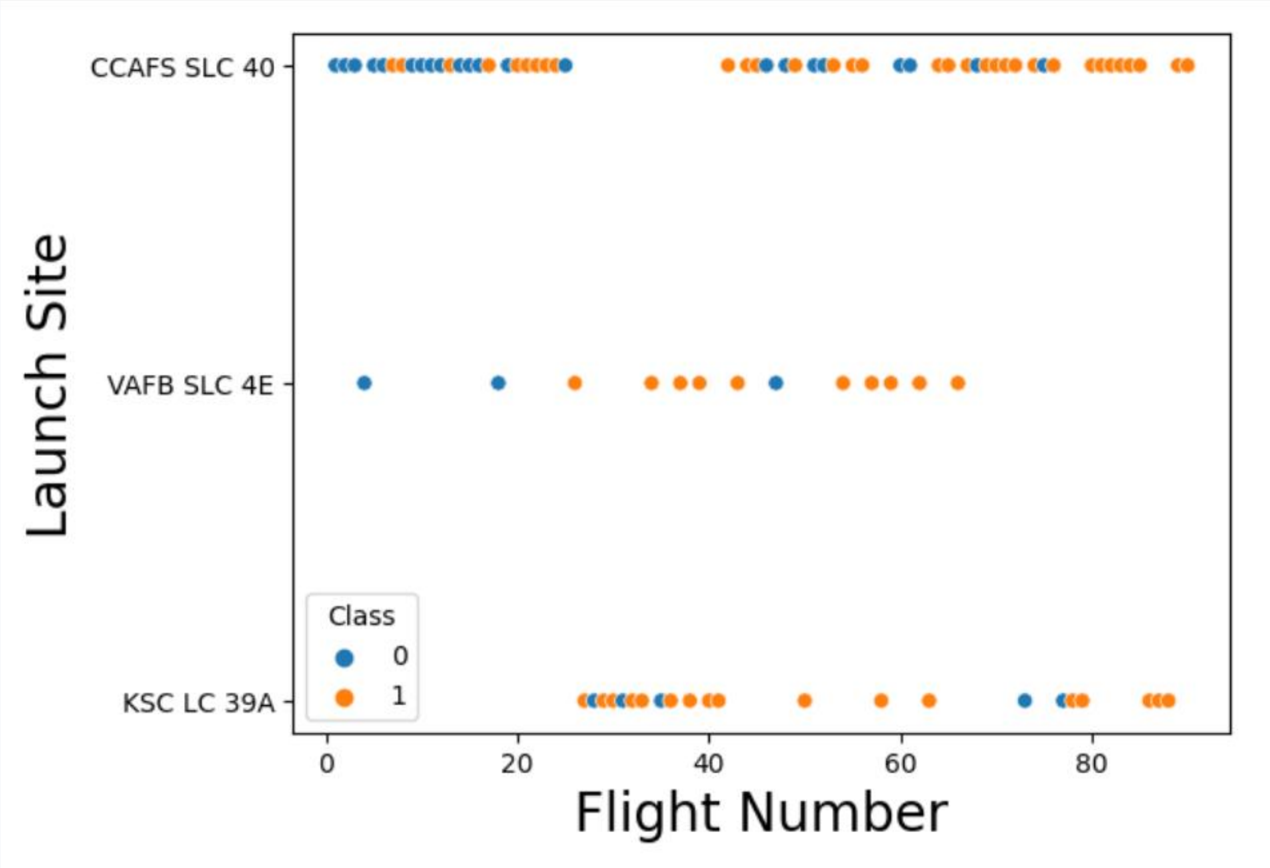
- Model accuracy scores showed all four models have 83% accuracy

Section 2

Insights drawn from EDA

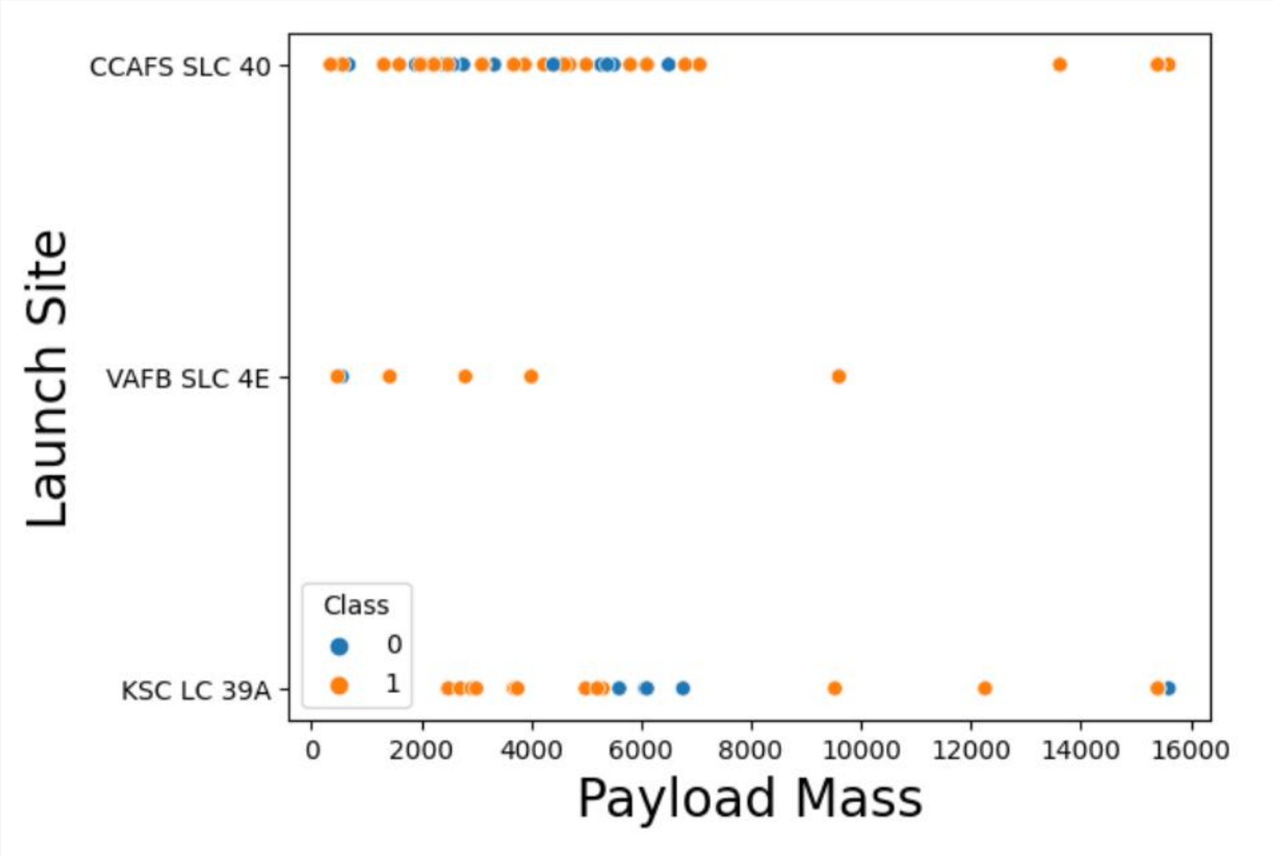
Flight Number vs. Launch Site

- Launch success increases as flight number increases, indicating with more flights the more experienced Space X got and success rate got better.



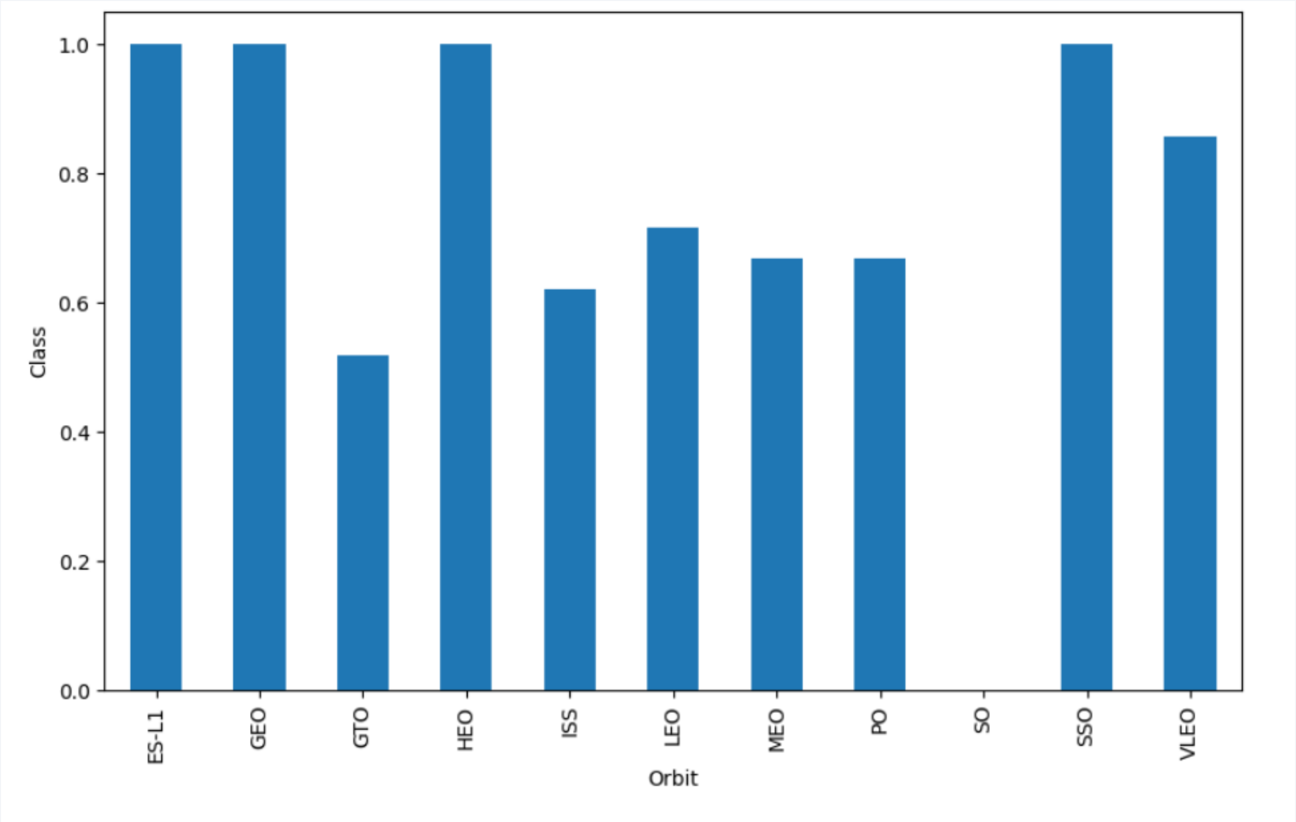
Payload vs. Launch Site

- Most unsuccessful outcomes occurred between payload mass of 4000-8000 kg
- VAFB SLC 4E didn't have any rockets launching greater than 10000 kg payload mass



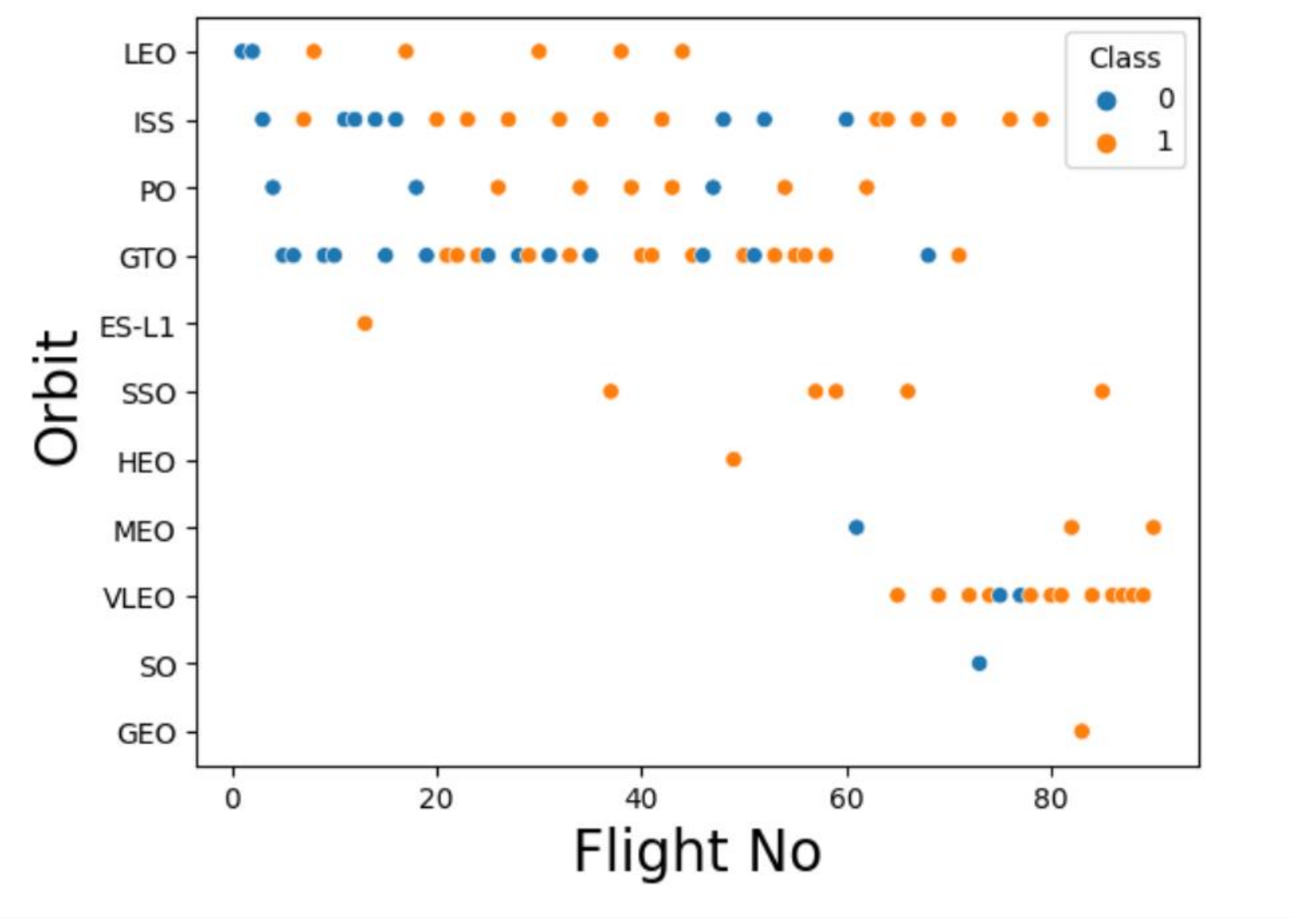
Success Rate vs. Orbit Type

- Success rate of 4 orbits (ES-L1, HEO, GEO, SSO) was 100%, making them highest successful. Lowest success rate orbits were SO (0%), and GTO (50%)



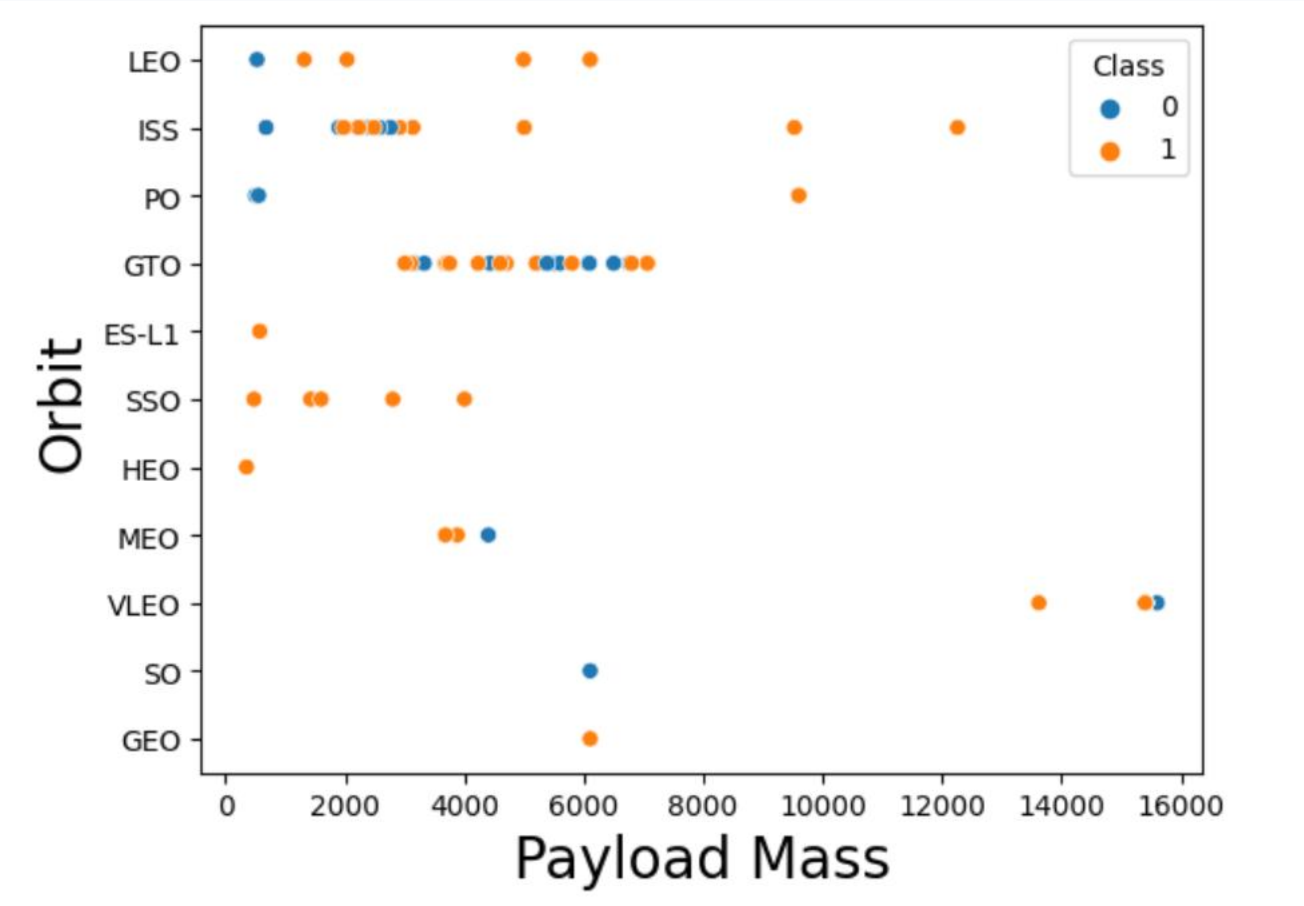
Flight Number vs. Orbit Type

- Certain orbits haven't been used past a certain flight no (e.g. LEO after flight 50). LEO orbit success correlated with flight number, which wasn't the case in VLEO. Orbits like ISS and GTO have been used for a long time period.



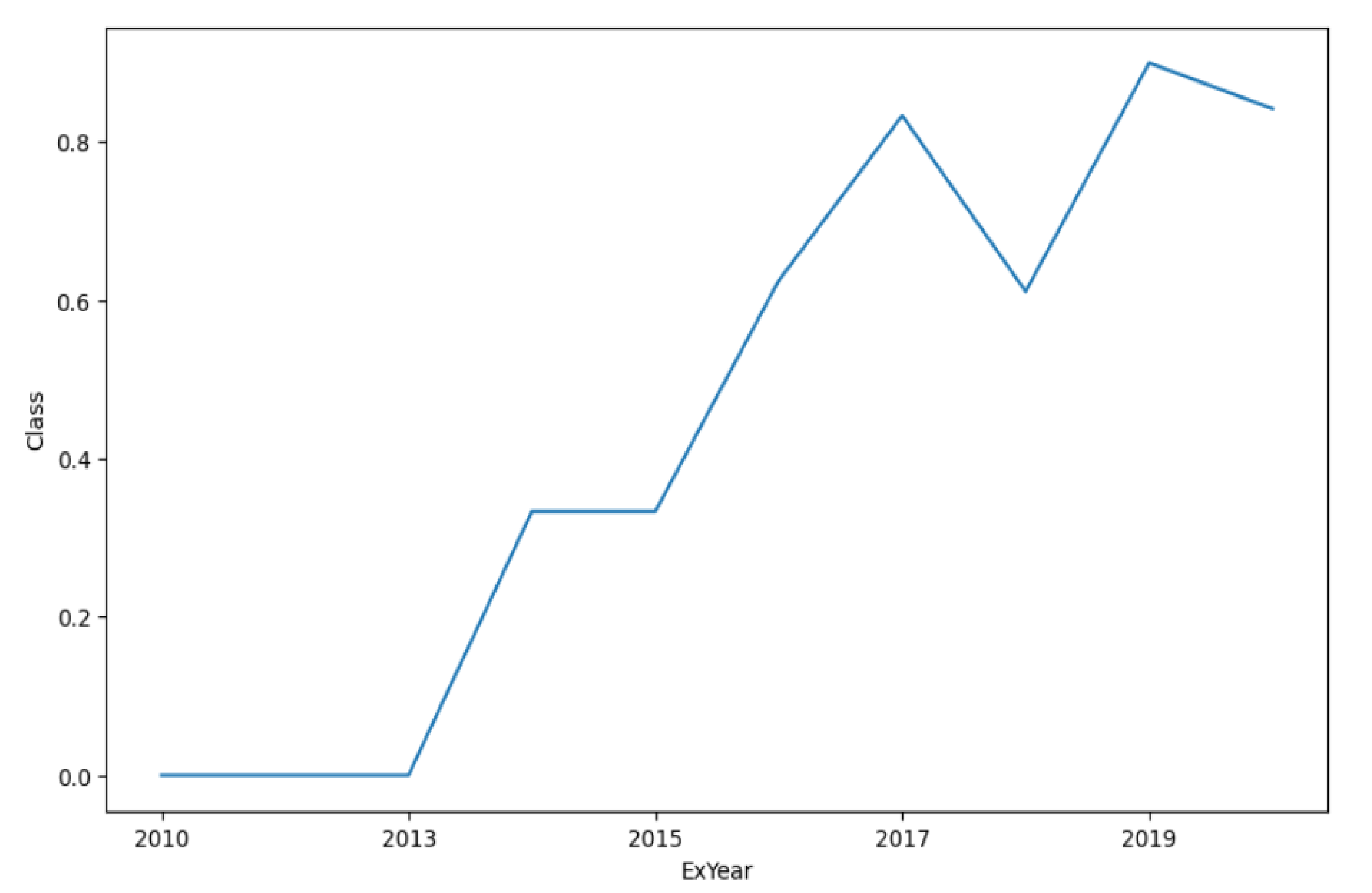
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are higher for Polar, LEO and ISS.



Launch Success Yearly Trend

- Success rate since 2013 kept increasing till 2020



All Launch Site Names

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

```
%%sql
select distinct Launch_site
from spacextbl
```

Pytl

```
* sqlite:///my_data1.db
Done.
```

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

Launch Site Names Begin with 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
select * from spacextbl where Launch_Site like 'CCA%' limit 5
```

Python

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Cu
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	
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)	
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	

Total Payload Mass

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
select sum(PAYLOAD_MASS__KG_) from spacextbl where Customer = 'NASA (CRS)'
```

Python

```
 * sqlite:///my_data1.db
Done.
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

Display average payload mass carried by booster version F9 v1.1

```
%%sql
select avg(PAYLOAD_MASS__KG_)
from spacextbl
where Booster_Version like "F9 v1.1"
```

Python

```
* sqlite:///my_data1.db
Done.
```

avg(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql
select min(Date) from spacextbl
where `Landing _Outcome` like 'Success (ground pad)'
```

12]

Python

```
 * sqlite:///my_data1.db
Done.
```

min(Date)
01-05-2017

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
select distinct Booster_Version, Mission_Outcome, payload_mass__kg_
from spacextbl
where "mission_outcome" like "%success%"
and PAYLOAD_MASS__KG_ between 4000 and 6000
```

Booster_Version	Mission_Outcome	PAYLOAD_MASS__KG_
F9 v1.1	Success	4535
F9 v1.1 B1011	Success	4428
F9 v1.1 B1014	Success	4159
F9 v1.1 B1016	Success	4707
F9 FT B1020	Success	5271
F9 FT B1022	Success	4696
F9 FT B1026	Success	4600
F9 FT B1030	Success	5600
F9 FT B1021.2	Success	5300
F9 FT B1032.1	Success	5300
F9 B4 B1040.1	Success	4990
F9 FT B1031.2	Success	5200
F9 B4 B1043.1	Success (payload status unclear)	5000
F9 FT B1032.2	Success	4230
F9 B4 B1040.2	Success	5384
F9 B5 B1046.2	Success	5800
F9 B5 B1047.2	Success	5300
F9 B5 B1046.3	Success	4000
F9 B5B1054	Success	4400
F9 B5 B1048.3	Success	4850
F9 B5 B1051.2	Success	4200
F9 B5B1060.1	Success	4311
F9 B5 B1058.2	Success	5500
F9 B5B1062.1	Success	4311

Total Number of Successful and Failure Mission Outcomes

List the total number of successful and failure mission outcomes

```
[16]
%%sql

select count("mission_outcome")
from SPACEXTBL
where "mission_outcome" like "%success%" or
    "mission_outcome" like "%failure%";
```

```
 * sqlite:///my_data1.db
Done.
```

count("mission_outcome")
101

Boosters Carried Maximum Payload

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
select distinct Booster_Version,  payload_mass__kg_ from spacextbl
where PAYLOAD_MASS__KG_= (select max(PAYLOAD_MASS__KG_) from spacextbl)
```

Python

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

```
%%sql
select substr(Date, 4, 2) as month, Booster_Version, Launch_Site, `Landing _Outcome`
from spacextbl where `Landing _Outcome`
like 'Failure (drone ship)' and substr(Date,7,4)='2015'
```

Python

```
* sqlite:///my_data1.db
Done.
```

month	Booster_Version	Launch_Site	Landing _Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

```
%%sql
select count(`Landing _Outcome`), `Landing _Outcome` from spacextbl
where `Landing _Outcome` like 'Success%'
and Date between '04-06-2010' and '20-03-2017'
group by `Landing _Outcome` order by count(`Landing _Outcome`) desc
```

[19] Python

```
 * sqlite:///my_data1.db
Done.
```

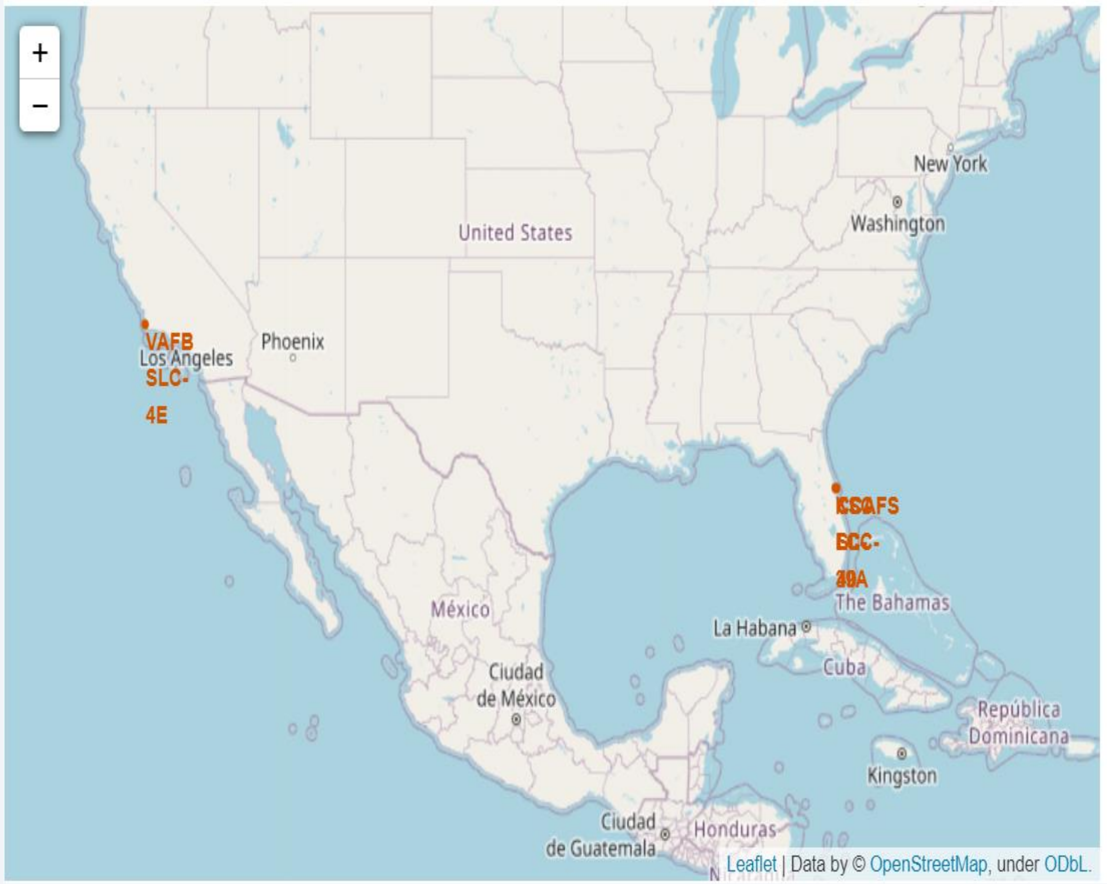
count(`Landing _Outcome`)	Landing _Outcome
20	Success
8	Success (drone ship)
6	Success (ground pad)

Section 3

Launch Sites Proximities Analysis

Locations of Each Launch Site

- The orange markers indicate the locations of each launch site, there are 4 launch sites with 3 of them on the east coast and one on the west coast.

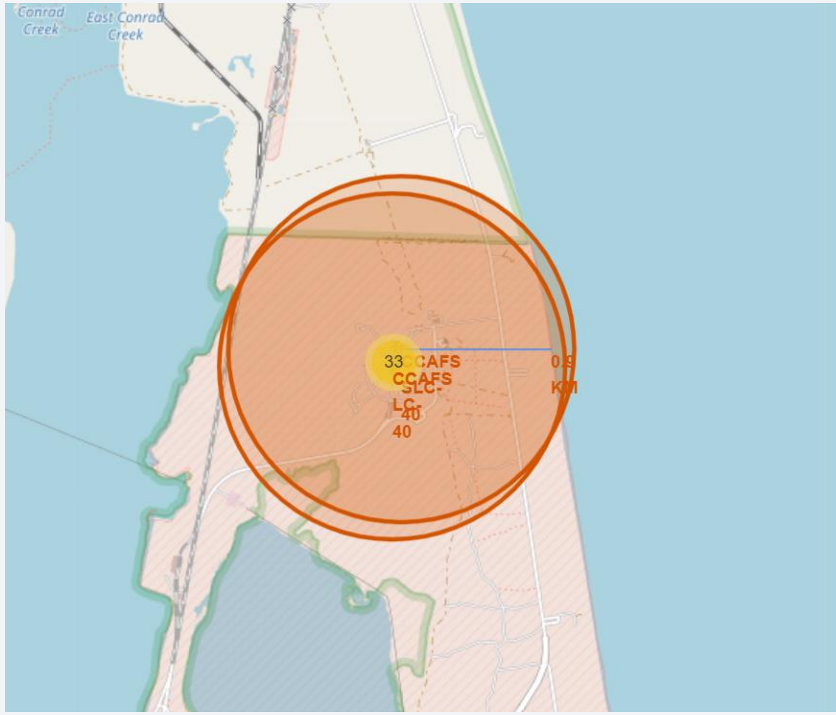


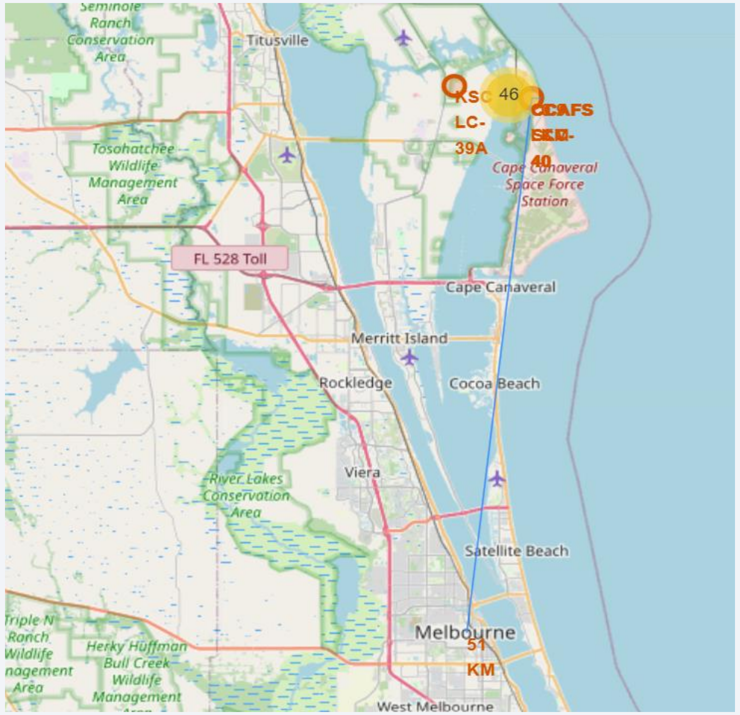
Launch Outcome for Each Site on the Map

- This map shows the launch outcomes on the map. When clicked on the cluster icons, namely the 10 and 46, green and red icons pop up that show success and failure outcomes respectively.

Distance to Proximities

- As seen on the screenshots, the launch site CCAFS SLC-40 is 0.9 km from the nearest coastline and 51 KM from the city of Melbourne.



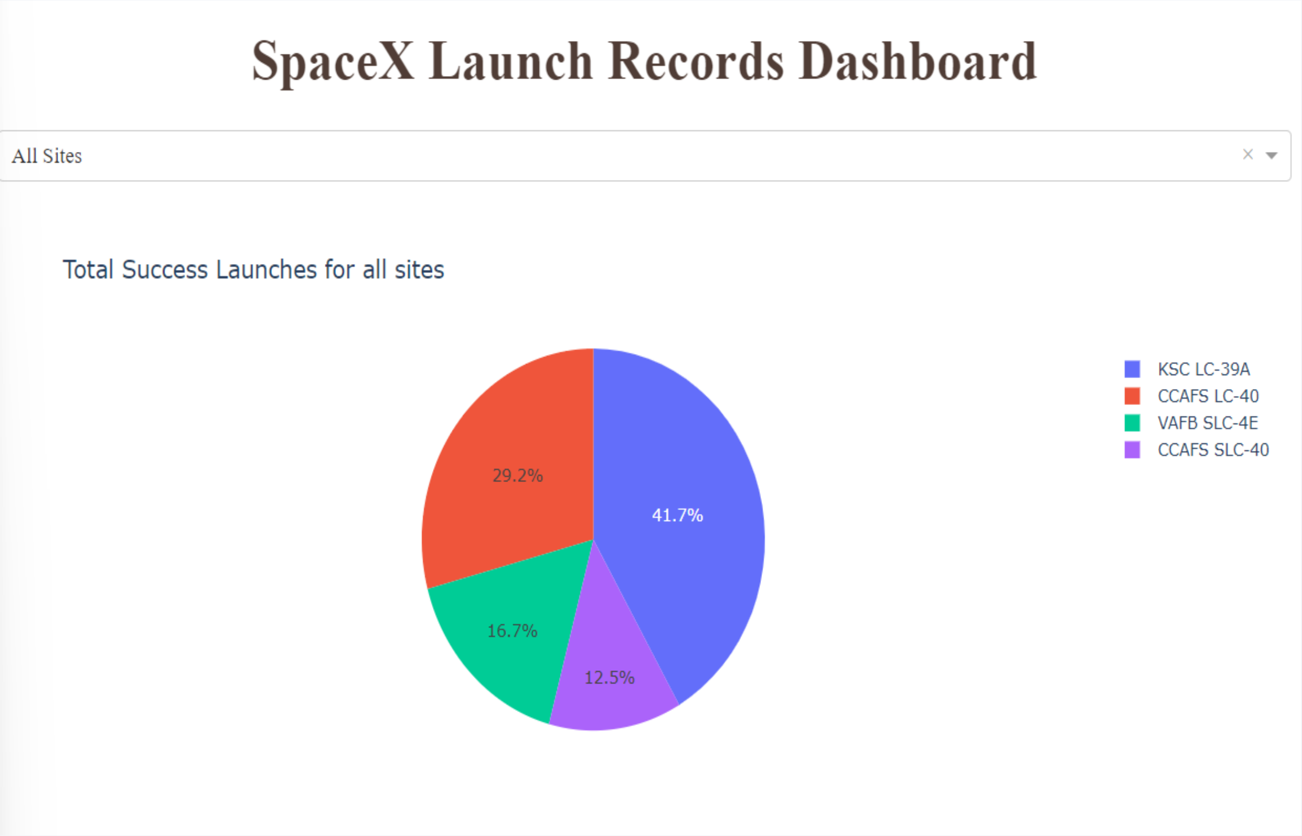


Section 4

Build a Dashboard with Plotly Dash

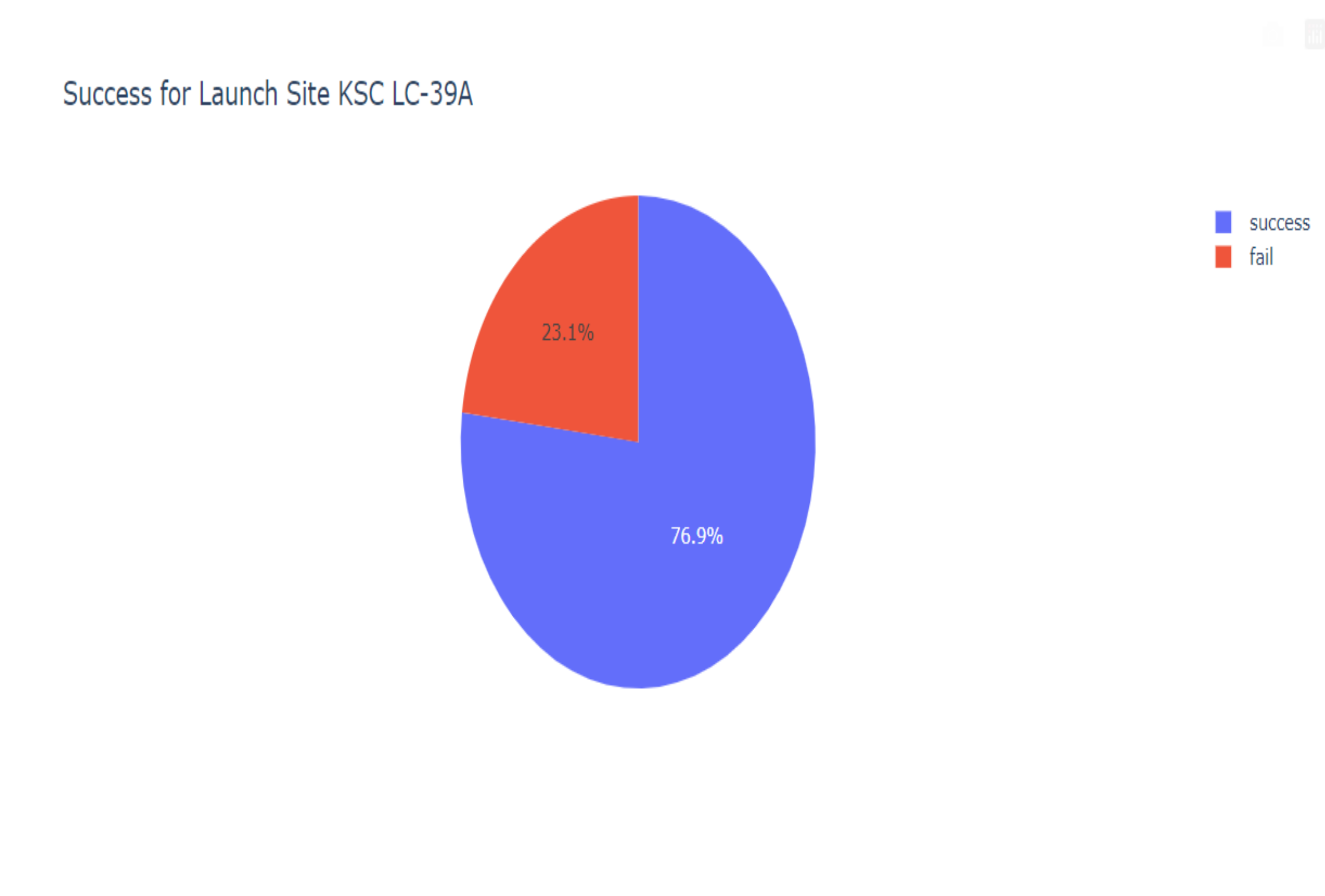
Launch Success Counts for All Sites

- This pie chart shows total success launches for all sites. KSC LC-39A amounts for the largest amount of total success launches at 41.7%.



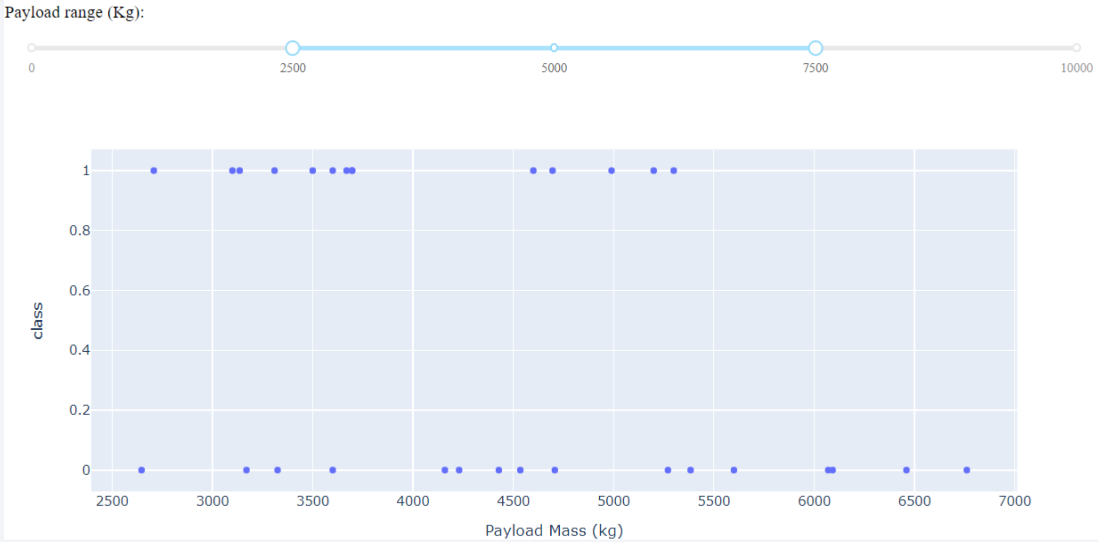
Highest Launch Success Ratio

- KSC LC-39A also had the highest launch success ratio at 76.9%



Payload vs Launch Outcome

- These plots show payload vs launch outcome graphs, with the first plot showing outcomes across all payload masses and the second showing outcomes between payload mass 2500 and 7500- as seen on the slider settings.
- These plots show that as the payload mass goes above 5.5K band, the success rate significantly decreases.

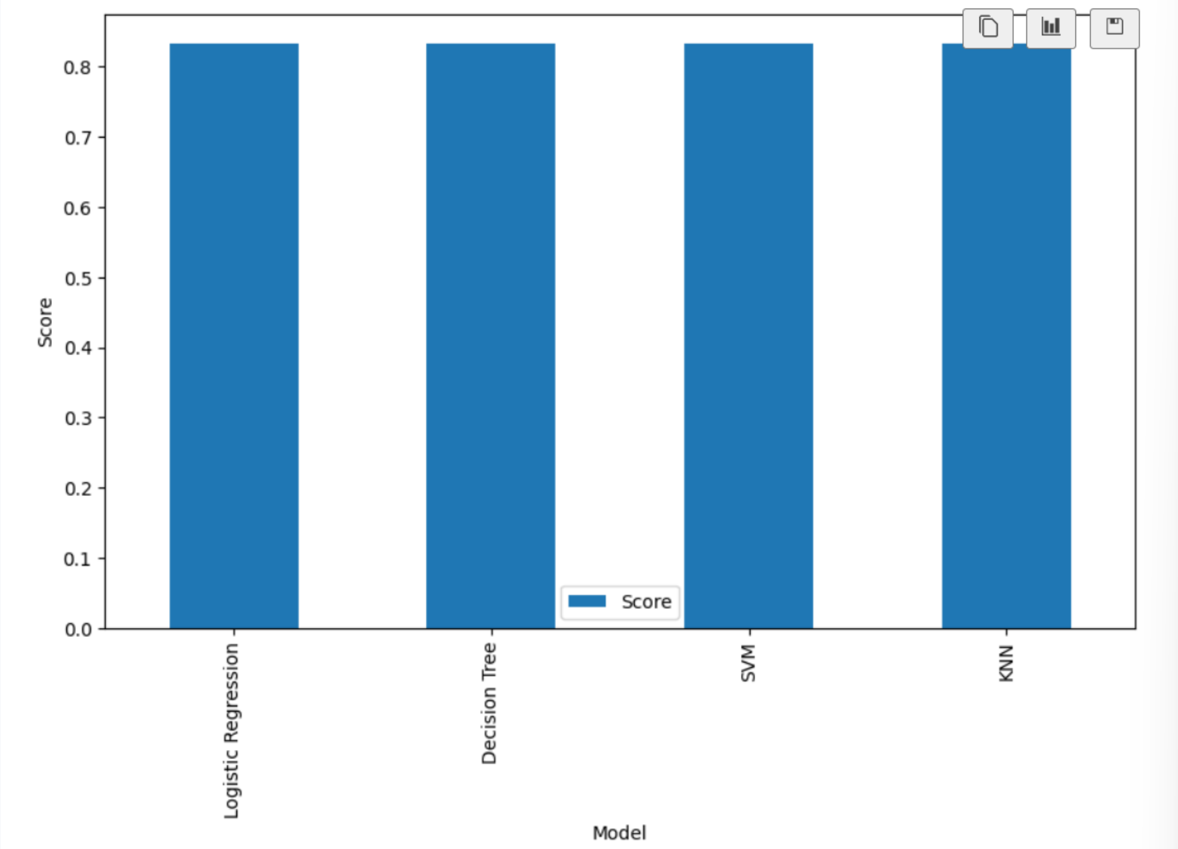


Section 5

Predictive Analysis (Classification)

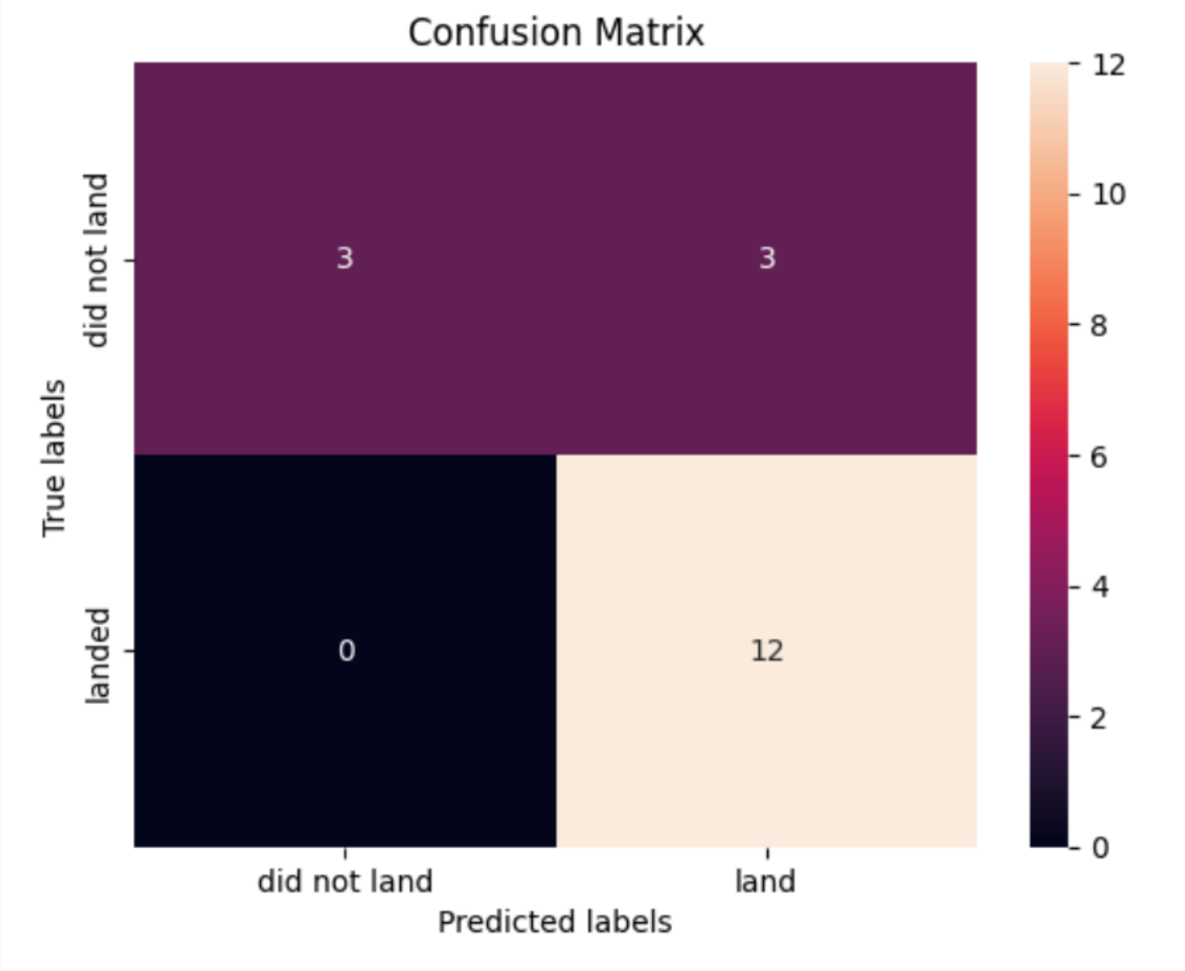
Classification Accuracy

- As seen on the bar chart, all models have yielded the same prediction accuracy (83.33%)



Confusion Matrix

- All models had the same confusion matrix as shown



Conclusions

- In this project, machine learning models were developed to predict the success of landing of the first stage of a rocket launch using data from Space X.
- 4 different machine learning models were trained for modeling: Logistic Regression, SVM, Decision Tree, KNN.
- The results have indicated that all four models yield the same prediction accuracy

Appendix

- Please visit https://github.com/mehmetbe/IMB-DS-Capstone for all codes and notebooks

Thank you!