IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Saba Mirsattari
15th of Jan 2023

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection through API
  - Data Collection with Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL
  - Exploratory Data Analysis with Data Visualization
  - Building an interactive map with Folium
  - Building a dashboard with Plotly Dash
  - Predictive analysis
- Summary of all results
  - Exploratory Data Analysis results
  - Interactive analytics demo in screenshots
  - Predictive Analytics result

Introduction

- Project background and context

  SpaceX is the most successful company of the commercial space age, making space travel affordable. The company 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. This information can be used if an alternate company wants to bid against space X for a rocket launch. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

- Problems you want to find answers

  - How do variables such as payload mass, launch site, number of flights or orbit affect the success of the first stage of landing?
  - The interaction amongst various features that determine the success rate of a successful landing.
  - What operating conditions needs to be in place to ensure a successful landing program.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Using SpaceX API
  - Using web scraping from Wikipedia.
- Perform data wrangling
  - Dealing with missing value and filtering the data
  - Using One Hot Encoding to prepare the data to a binary classification
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models to ensure the best results

Data Collection

The data was collected using SpaceX REST API and webscrapping Wikipedia

- The information obtained by the API are rocket , launches, payload information
- The SpaceX REST API URL is : https://api.spacexdata.com/v4/launches/past

- The information collected by webscrapping of Wikipedia are launches, landing and payload information

Data Collection - SpaceX API

1. **Getting Response from API**

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

2. **Convert Response to JSON file**

```
response = requests.get(static_json_url).json()
data = pd.json_normalize(response)
```

3. **Apply custom function to clean data**

```
getLaunchSite(data)  getBoosterVersion(data)
getPayloadData(data) getCoreData(data)
```

4. **Assign list to dictionary then dataframe**

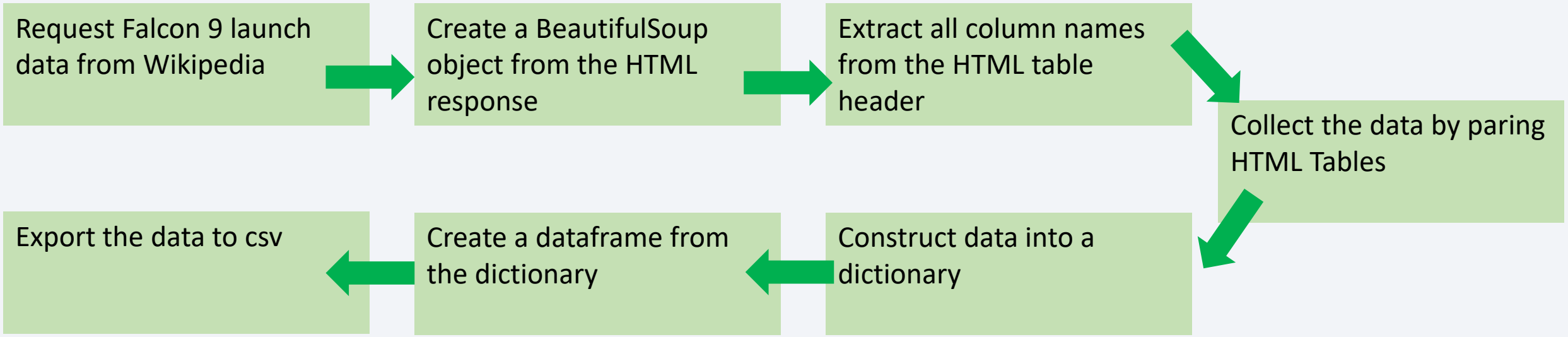
```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
df = pd.DataFrame.from_dict(launch_dict)
```

5. **Filter Dataframe and Export to csv file**

```
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```


GitHub URL here: https://github.com/sabacoursera/IBM-Capstone-Project/blob/master/data%20collection.ipynb

Data Collection - Scraping



GitHub URL: https://github.com/sabacoursera/IBM-Capstone-Project/blob/master/data%20scraping.ipynb

Data Wrangling

- In the data set, there are several different cases where the booster did not land successfully:
  - True Ocean, True RTLS and True ADS mean the mission has been successful.
  - False Ocean, False RTLS , False ASDS mean the mission was a failure.

- We performed multiple exploratory data analyses and determined training labels:
  - Calculate the number of launches on each site
  - Calculate the number and occurrence of each orbit
  - Calculate the number of mission outcome per orbit type
  - Create a landing outcome label from Outcome column and export to csv:

Github URL : https://github.com/sabacoursera/IBM-Capstone-Project/blob/master/data%20wrangling.ipynb

EDA with Data Visualization

We explored the data by visualizing the relationship between variables using different chart types:

- Scatter plot : flight number vs payload mass, flight number vs launch site, payload mass vs launch site, flight number vs orbit type, payload mass vs orbit type.
- Bart chart: success rate of each orbit type
- Line chart : the launch success yearly trend.

Github URL: https://github.com/sabacoursera/IBM-Capstone-Project/blob/master/Exploratory%20using%20pandas%20and%20marplot.ipynb

EDA with SQL

- We loaded the SpaceX dataset into a SQL database and connectd to the jupyter notebook to run queries such as :
  - The names of 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
  - The total number of successful and failure mission outcomes
  - The failed landing outcomes in drone ship, their booster version and launch site names.

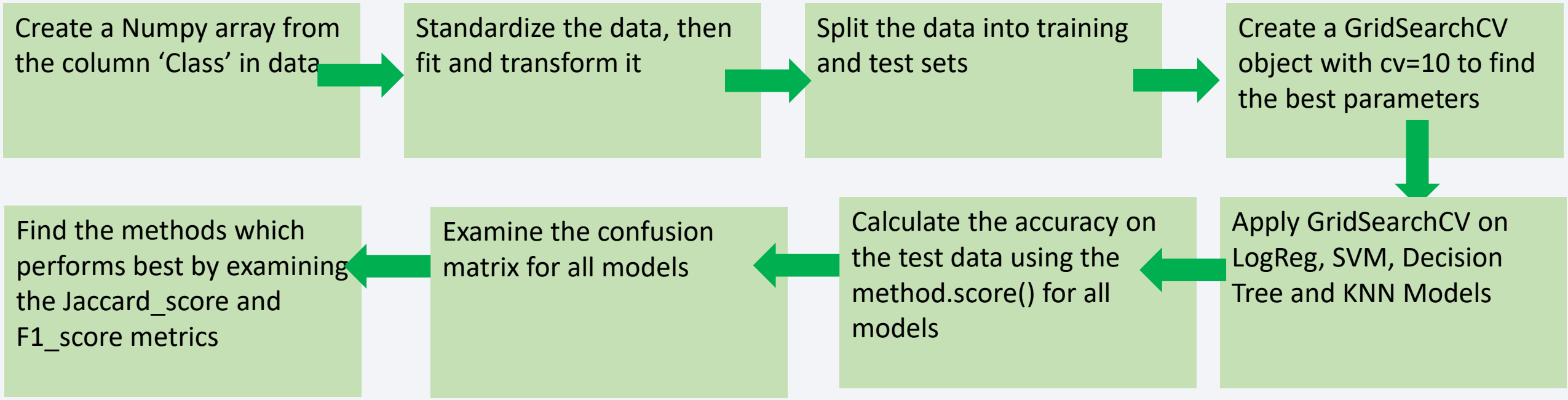
Github URL: https://github.com/sabacoursera/IBM-Capstone-Project/blob/master/exploratory%20using%20sql.ipynb

Build an Interactive Map with Folium

- Markers, circles, line and marker clusters were used with Folium maps
  - Markets indicate points like launch sites
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Center
  - Market clusters indicate groups of evets in each coordinate, like launches in a launch site
  - Lines are used to indicate distance between 2 coordinates

Github URL: https://github.com/sabacoursera/IBM-Capstone-Project/blob/master/Visual%20Analytics%20with%20Folium%20lab.ipynb

Predictive Analysis (Classification)



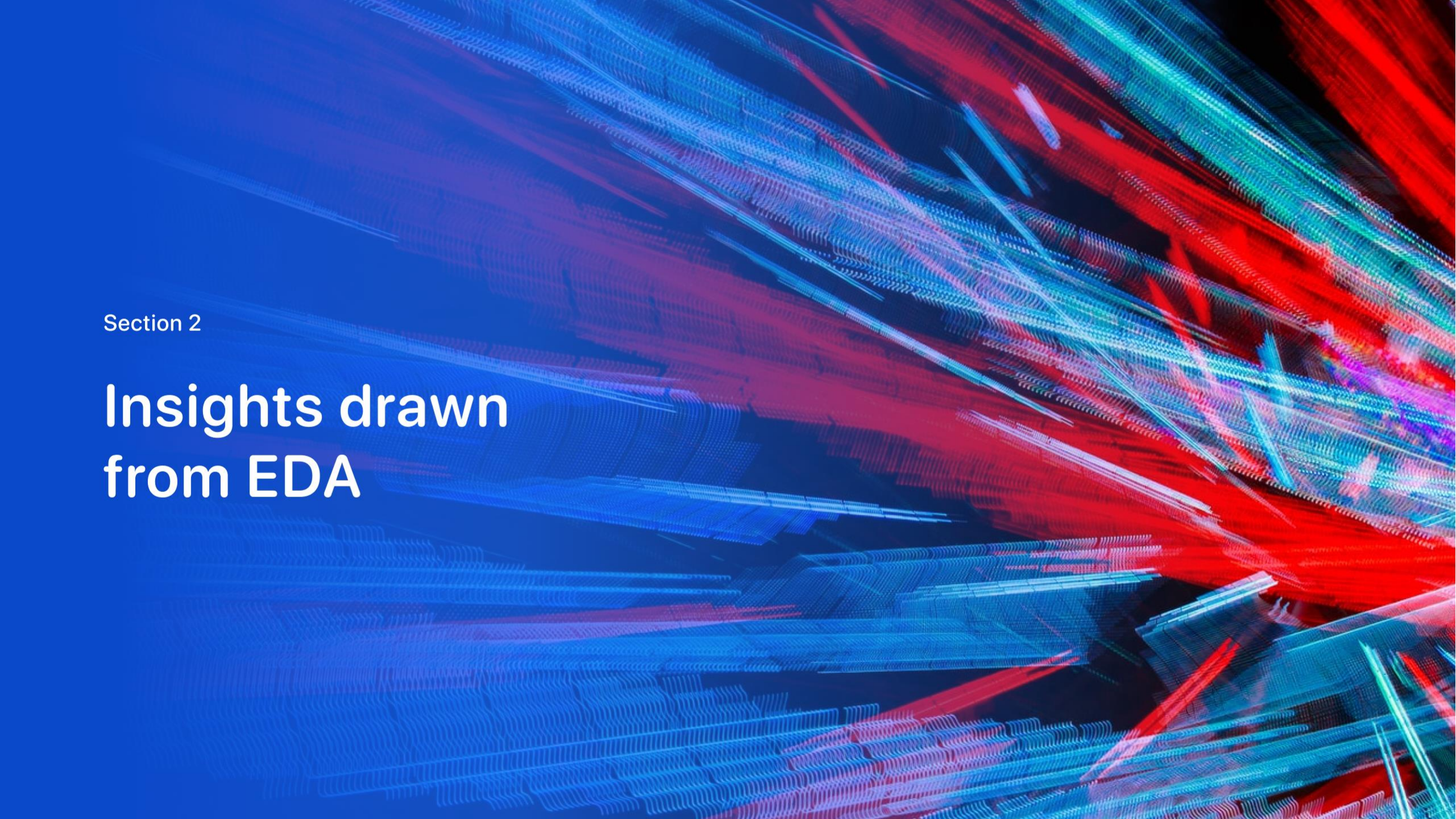
GitHub URL: https://github.com/sabacoursera/IBM-Capstone-Project/blob/master/machine%20learning%20predictions.ipynb

Results

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

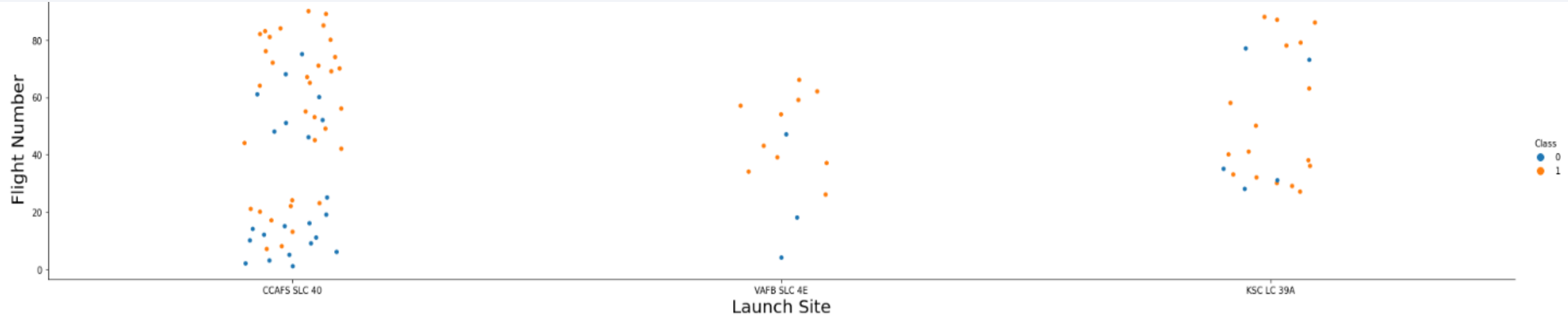
Section 2

Insights drawn from EDA



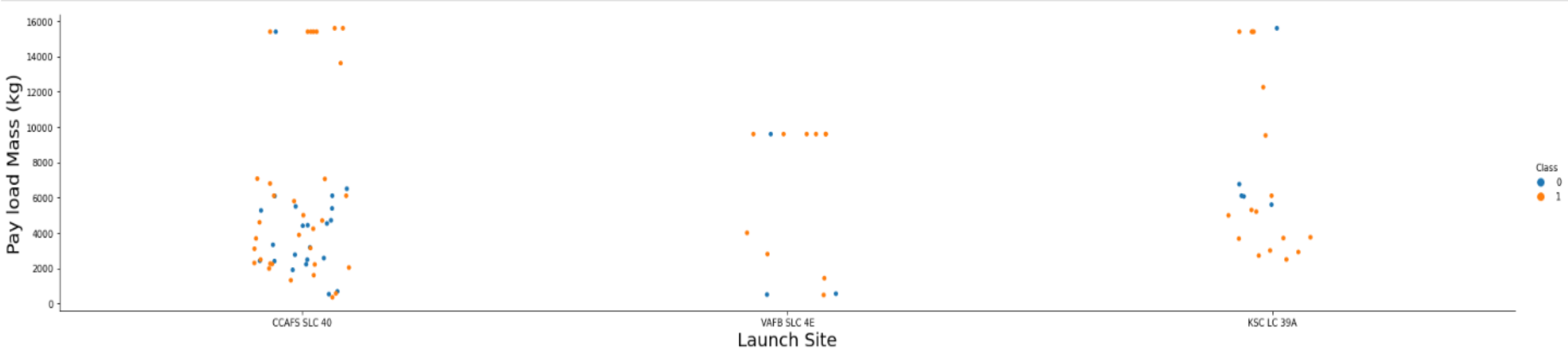
Flight Number vs. Launch Site

- From the plot, we found that the earliest flights all failed while the latest flight all succeeded
- KSC LC-39A and VAFB SLC 4E have higher success rates



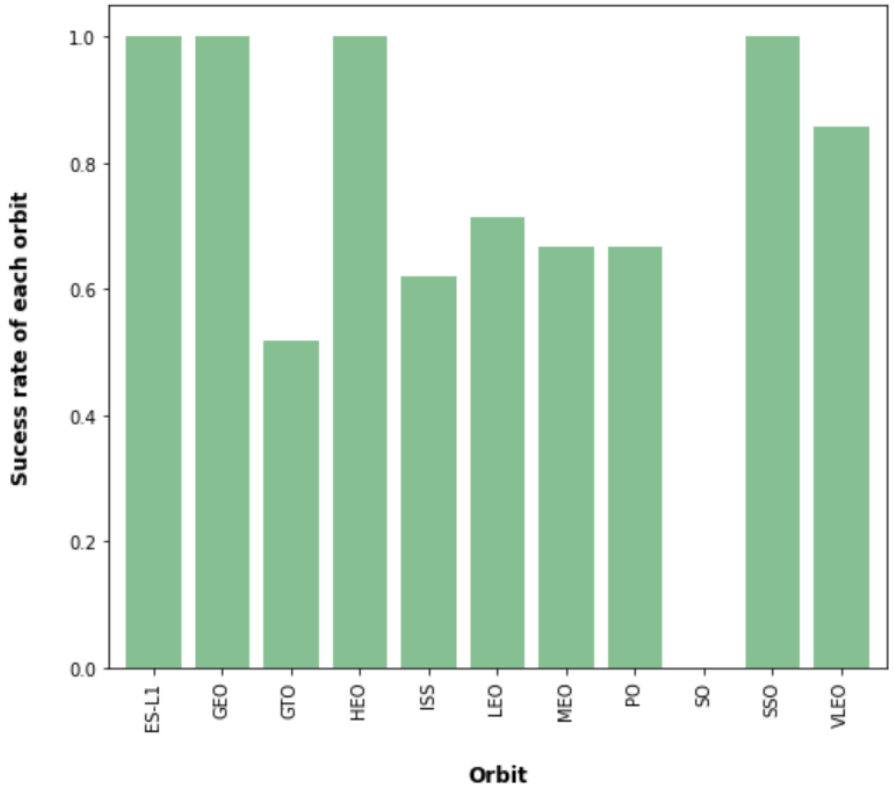
Payload vs. Launch Site

- For the VAFB-SLC launch site there are no rockets launched for heavy payload mass, greater than 10000 kg.



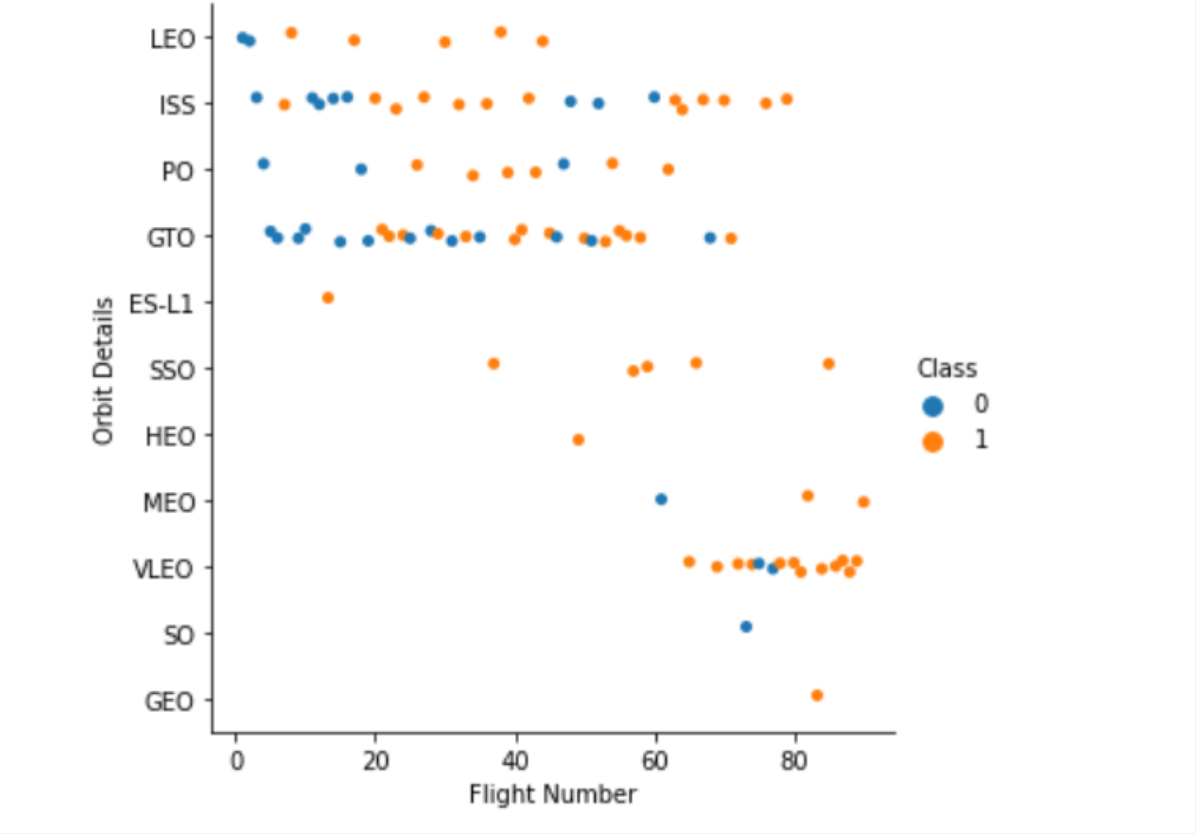
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, SSO has highest Success rates. SO has poorest.



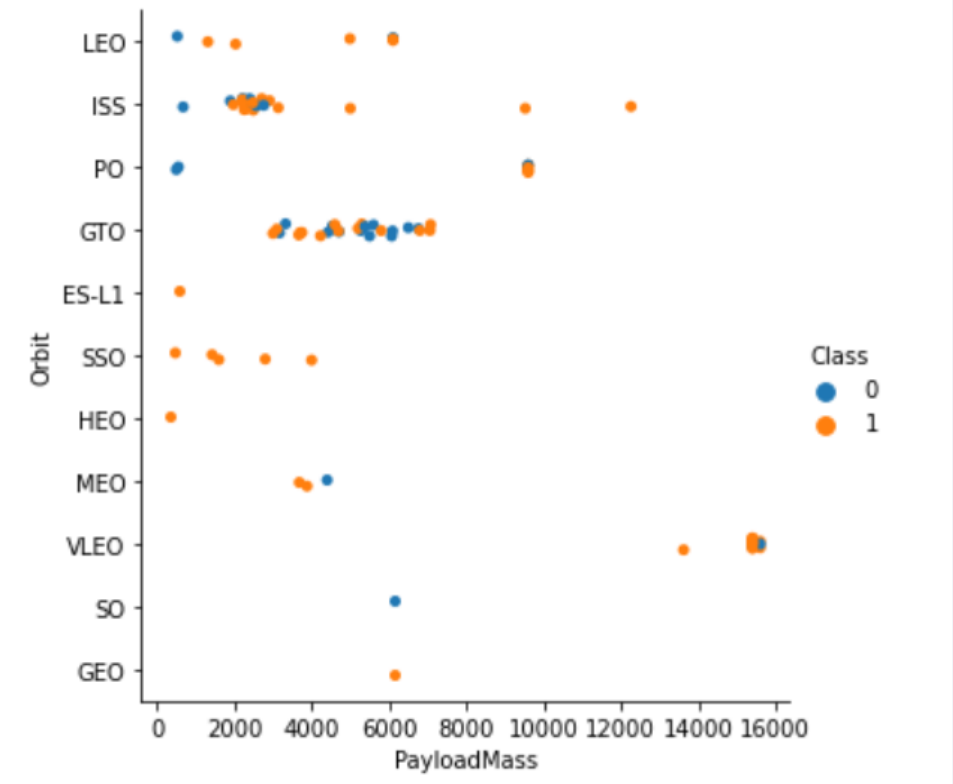
Flight Number vs. Orbit Type

In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



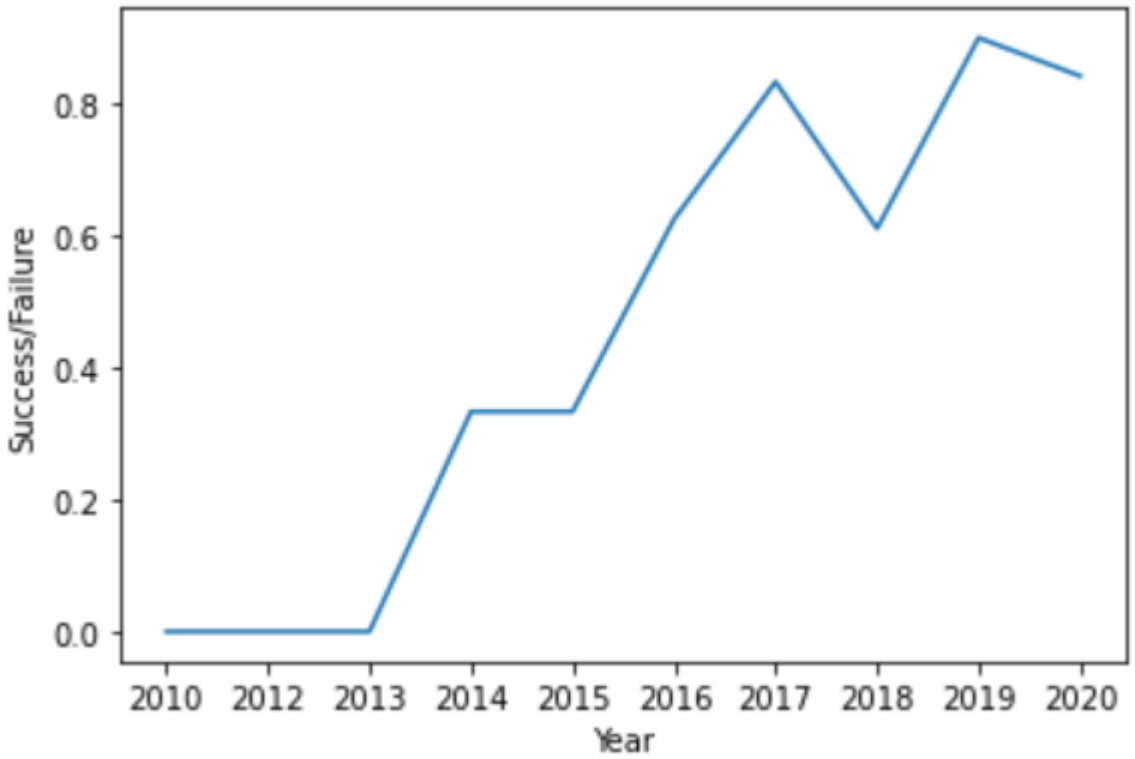
Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits. However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020.



All Launch Site Names

- We used the SQL Query to select distinct launch sites from SPACEXTBL to show only unique launch sites from the SpaceX data.

```
%sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEXTBL;
```

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

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

Done.

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

- We used the query above to display 5 records where launch sites begin with `CCA`

Total Payload Mass

- We calculated the total payload carried by boosters from NASA as 45596 using the query below

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass by NASA (CRS)" FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)';
```

Done.

Total Payload Mass by NASA (CRS)
45596

Average Payload Mass by F9 v1.1

- We calculated the average payload mass carried by booster version F9 v1.1 as 2928.4

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload Mass by Booster Version F9 v1.1" FROM SPACEXTBL \
WHERE BOOSTER_VERSION = 'F9 v1.1';
```

Done.

Average Payload Mass by Booster Version F9 v1.1
2928

First Successful Ground Landing Date

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015

```
%sql SELECT MIN(DATE) AS "First Succesful Landing Outcome in Ground Pad" FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

First Succesful Landing Outcome in Ground Pad
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

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

Total Number of Successful and Failure Mission Outcomes

- We used wildcard like ‘%’ to filter for **WHERE** MissionOutcome was a success or a failure.

```
%sql SELECT COUNT(MISSION_OUTCOME) AS "Successful Mission" FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'Success%';
```

Successful Mission
100

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the **WHERE** clause and the **MAX()** function.

```
%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEXTBL \
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

Booster Versions which carried the Maximum Payload Mass
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

2015 Launch Records

- We used a combinations of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015

```
%sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE DATE LIKE '2015-%' AND \
LANDING__OUTCOME = 'Failure (drone ship)';
```

b

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

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

- We selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 2010-06-04 to 2010-03-20.
- We applied the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.

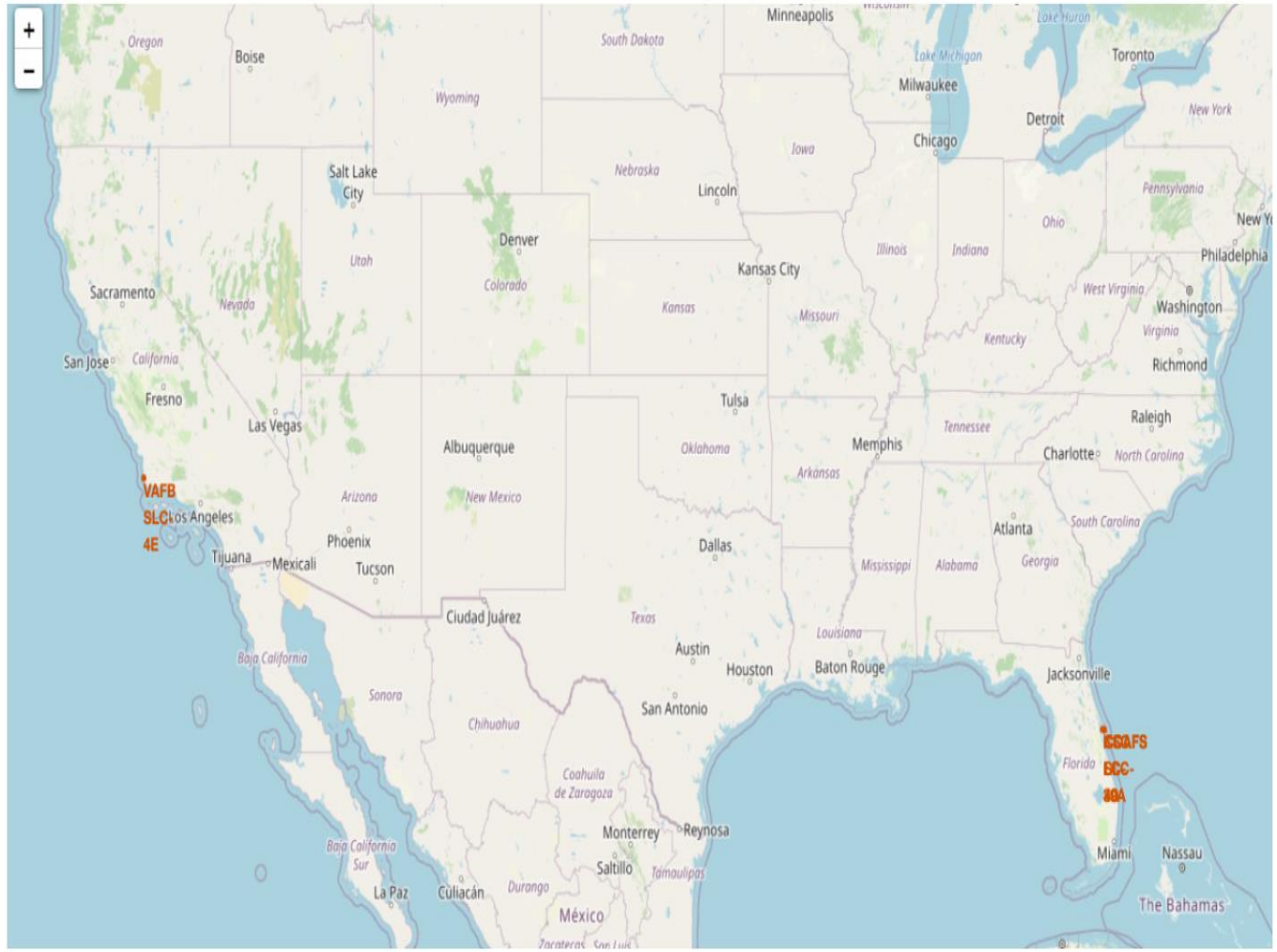
```
%sql SELECT LANDING__OUTCOME as "Landing Outcome", COUNT(LANDING__OUTCOME) AS "Total Count" FROM SPACEXTBL \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC ;
```

Landing Outcome	Total 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

Section 4

Launch Sites Proximities Analysis

All launch sites global map markers



All launch sites are in very close proximity to the coast. The rocket would have the advantage of flying over the ocean, minimizing the risk of having any debris dropping or exploding near people.

Most of launch sites are in close proximity to the Equator line. If a spacecraft is launched from a site near Earth's equator, it can take optimum advantage of the Earth's substantial rotational speed. Sitting on the launch pad near the equator, it is already moving at a speed of over 1650 km per hour relative to Earth's center.

Color-labeled launch records on the map

Explanation:

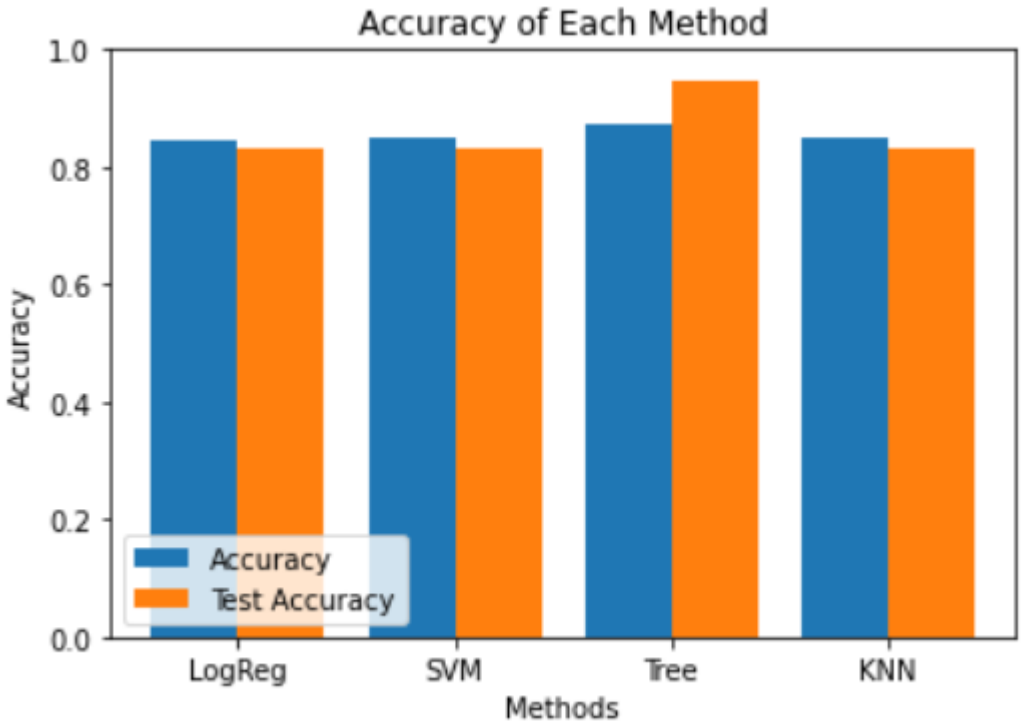
- We created markers for all launch records. If a launch was successful (class=1), then we used a green marker and if a launch was failed, we use a red marker (class=0)
-

Section 6

Predictive Analysis (Classification)

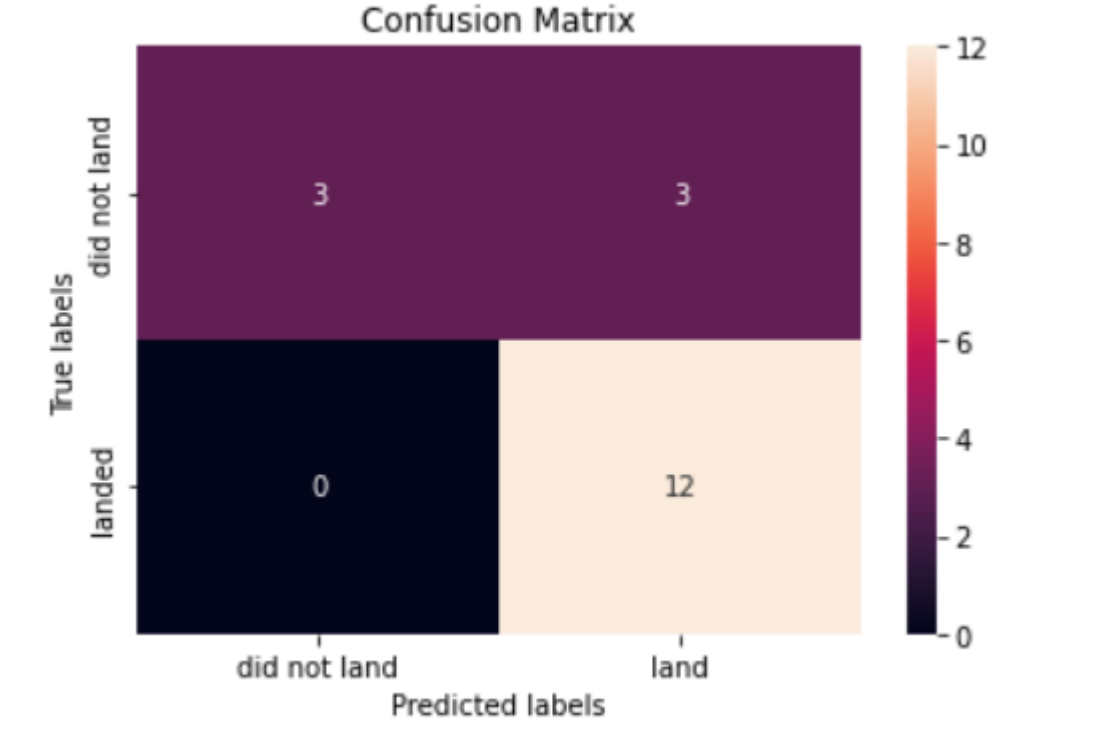
Classification Accuracy

Based on the scores of the Test Set, we can not confirm which method performs best. Same Test Set scores may be due to the small test sample size (18 samples) . Therefore, we tested all methods based on the whole Dataset. The scores of the whole Dataset confirm that the best model is the Decision Tree Model.



Confusion Matrix

- The confusion matrix for the decision tree classifier shows that logistic regression can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best algorithm for this dataset
- Most of the launch sites are in proximity of the Equator line and all the sites are in very close proximity to the coast

Thank you!