IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

CLAUDIO SAN MARTIN
2024-02-22

Outline

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

Executive Summary

- We have collected information through the SpaceX API and scraping Wikipedia, since it contains a lot of detailed information about it, which has made our work easier.
- The data has been analyzed with techniques such as EDA and visual analytics with the Folium library.
- Through the analysis carried out on the SpaceX Falcon-9 launches we have been able to obtain some interesting perspectives, which, applied through different methodologies, we have been able to create a statistical model which allows us to predict the result with 80% accuracy. of the upcoming launches and reuse of Falcon-9.

Introduction

SpaceX 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 SpaceX 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 SpaceX for a rocket launch.

This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

The data of SpaceX was gathered from the SpaceX REST API, giving us data about launches, rocket used, landing specifications, and landing outcome.

- Perform data wrangling

Trough EDA we will find patterns in the data and determine what would be the label for training.

- Perform exploratory data analysis (EDA) using visualization and SQL

Using MatPlotLib and Seaborn we will make visualizations that give us different perspectives on the data.

- Perform interactive visual analytics using Folium and Plotly Dash

We will be able to perform interactive analyzes

- Perform predictive analysis using classification models

Applying GridSearchCV we will find the best hyperparameters for SVM, Classification Trees and Logistic Regression, and select the method performs best.

Data Collection

The data of SpaceX was gathered from the SpaceX REST API, giving us data about launches, rocket used, landing specifications, and landing outcome.

We will use the API targeting endpoints to gather specific data for each ID number.

These functions are already created for you, and will use the following: Booster, Launchpad, payload, and core.

The data will be stored in lists and will be used to create our dataset.

We will be using the Python BeautifulSoup package to web scrape some HTML tables that contain valuable Falcon 9 launch records.

Data Collection – SpaceX API

GITHUB NOTEBOOK LINK

https://github.com/csanmartinabalos/SpaceX-Falcon9/blob/7335d1cd42785878c90a7e70d3342d7fccbd3e81/01.jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Scraping

WIKIPEDIA

WEB SCRAPING WITH BEAUTIFUL SOUP

DATASET

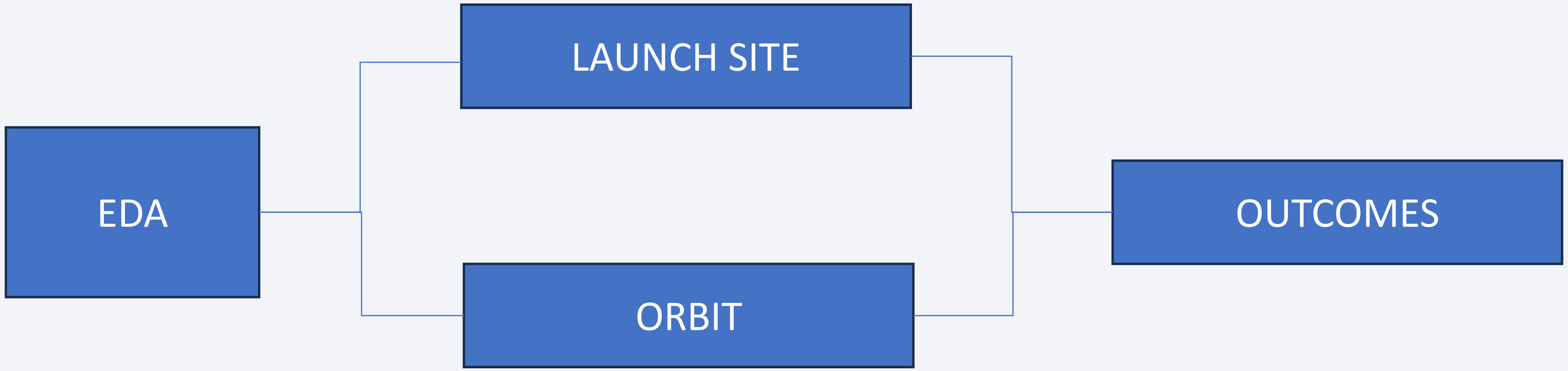
	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin1A	167.743129	9.047721
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2A	167.743129	9.047721
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin2C	167.743129	9.047721
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0	Merlin3C	167.743129	9.047721
4	6	2010-06-04	Falcon 9	NaN	LEO	CCAFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857

https://github.com/csanmartinabalos/SpaceX-Falcon9/blob/987a9c9ba8e8f46c2a33531644e0dc16a23228c2/02.jupyter-labs-webscraping.ipynb

Data Wrangling

There are several different cases where the booster did not land successfully. Sometimes a landing was attempted but failed due to an accident; for example, True Ocean means the mission outcome was successfully landed to a specific region of the ocean while False Ocean means the mission outcome was unsuccessfully landed to a specific region of the ocean.

We will mainly convert those outcomes into Training Labels with 1 means the booster successfully landed 0 means it was unsuccessful.



https://github.com/csanmartinabalos/SpaceX-Falcon9/blob/92847bae547b99cb5f43c8366f94689fbd4e7d33/03.labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb

EDA with Data Visualization

List of graphics used

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload Mass vs. Launch Site
- Orbit vs. Success rate
- Flight Number vs. Orbit
- Payload Mass vs. Orbit
- Year vs. Success rate

Thanks to the perspectives these graphs give us, we can say that SpaceX's success rate increases year after year and that Orbit and PayloadMass are fundamental factors in its performance.

https://github.com/csanmartinabalos/SpaceX-Falcon9/blob/3ec566a4bddab3a8e8c3d3f3bf2565c08cb2941d/04.jupyter-labs-eda-dataviz.ipynb

EDA with SQL

Queries used:

- SELECT DISTINCT Launch_site FROM SPACEXTABLE;
- SELECT * FROM SPACEXTABLE WHERE Launch_site LIKE 'KSC%’ LIMIT 5;
- SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass FROM SPACEXTABLE WHERE Customer LIKE '%NASA (CRS)%’
- SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1’;
- SELECT * FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (drone ship)’;
- SELECT * FROM SPACEXTABLE WHERE Landing_Outcome = 'Success (ground pad)’ AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
- SELECT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Success%' OR Mission_Outcome LIKE 'Failure%’ GROUP BY Mission_Outcome;
- SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);
- SELECT substr(Date, 6, 2) AS Month, Mission_Outcome AS Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(Date, 0, 5) = '2017' AND Landing_Outcome = 'Success (ground pad)’;
- SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03-20’ AND Landing_Outcome IN ('Failure (drone ship)', 'Success (ground pad)’) GROUP BY Landing_Outcome ORDER BY Outcome_Count DESC;

https://github.com/csanmartinabalos/SpaceX-Falcon9/blob/7b885a52cdad495c1aa7cb2b71f3a5759dd633ec/05.jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium

We use markers to locate all the launch points. In turn, we show how many launches have been successful and how many have failed.

Finally, we calculate the distance between the launch points and their proximities, to understand their location.

https://github.com/csanmartinabalos/SpaceX-Falcon9/blob/9d8514d8836e9a6fc76d8b61bcd739de84eaccd5/06.lab_jupyter_launch_site_location.jupyterlite.ipynb

Predictive Analysis (Classification)

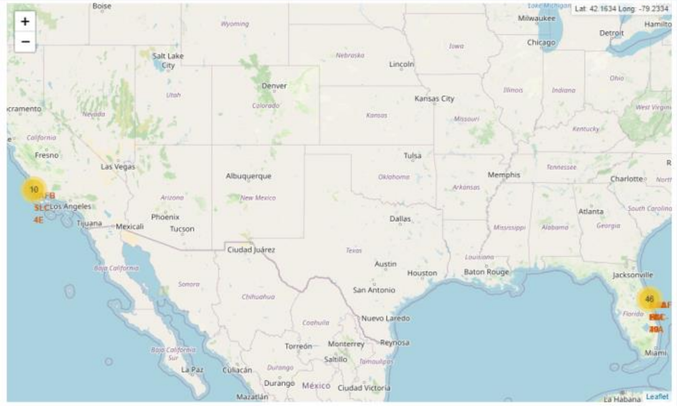
Once the final dataset has been obtained, it has been divided into training and test sets.

These have been preprocessed for standardization and several predictive models have been created to see which best fits the data.We have used Logistic Regression, SVM, Decision Tree and KNeighbors where we have selected the best hyperparameters of each one through GridSearchCV.

https://github.com/csanmartinabalos/SpaceX-Falcon9/blob/7bce1e649b2db6f09f044e851d8f8c4d59ffa2e8/07.SpaceX%20Project%20-%20Machine%20Learning%20Prediction.ipynb

Results

- We have been able to see that both Payload Mass and Orbit are fundamental elements in the success of the launches.
- Thanks to the interactive map visualization, we have been able to see that some launch points have a better success rate.

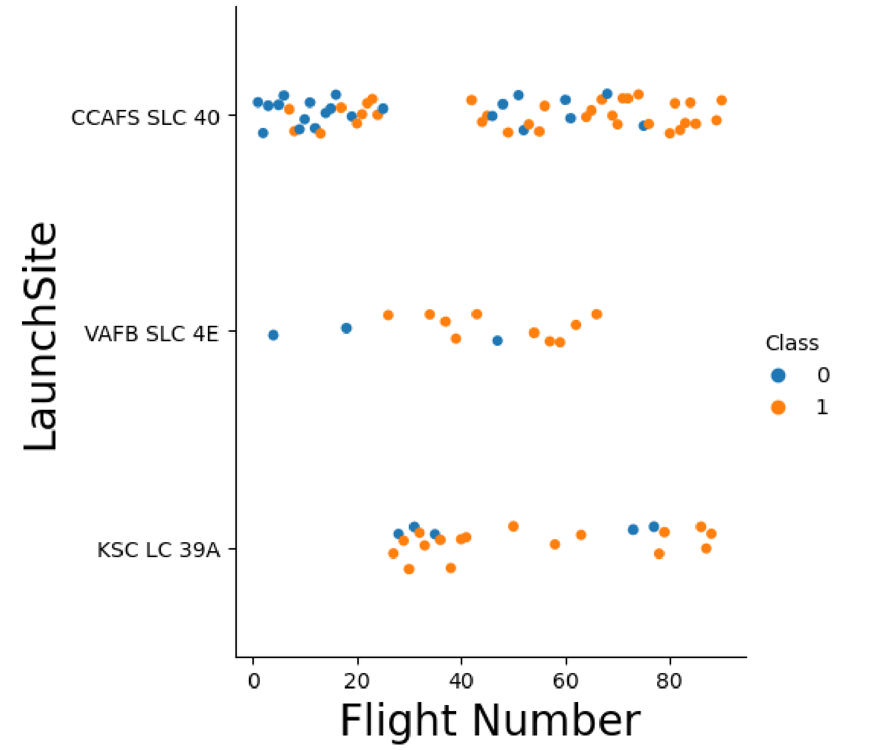


- Through machine learning we have been able to create a model which can predict with more than 80% accuracy whether the next launch will be successful or not.

Section 2

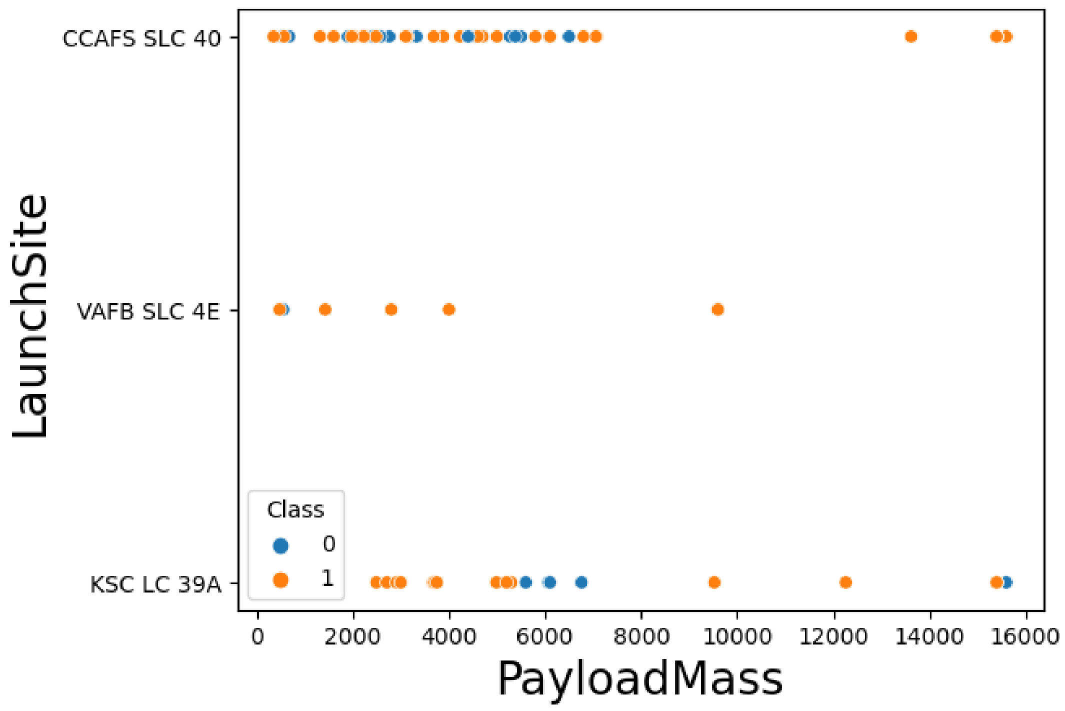
Insights drawn from EDA

Flight Number vs. Launch Site



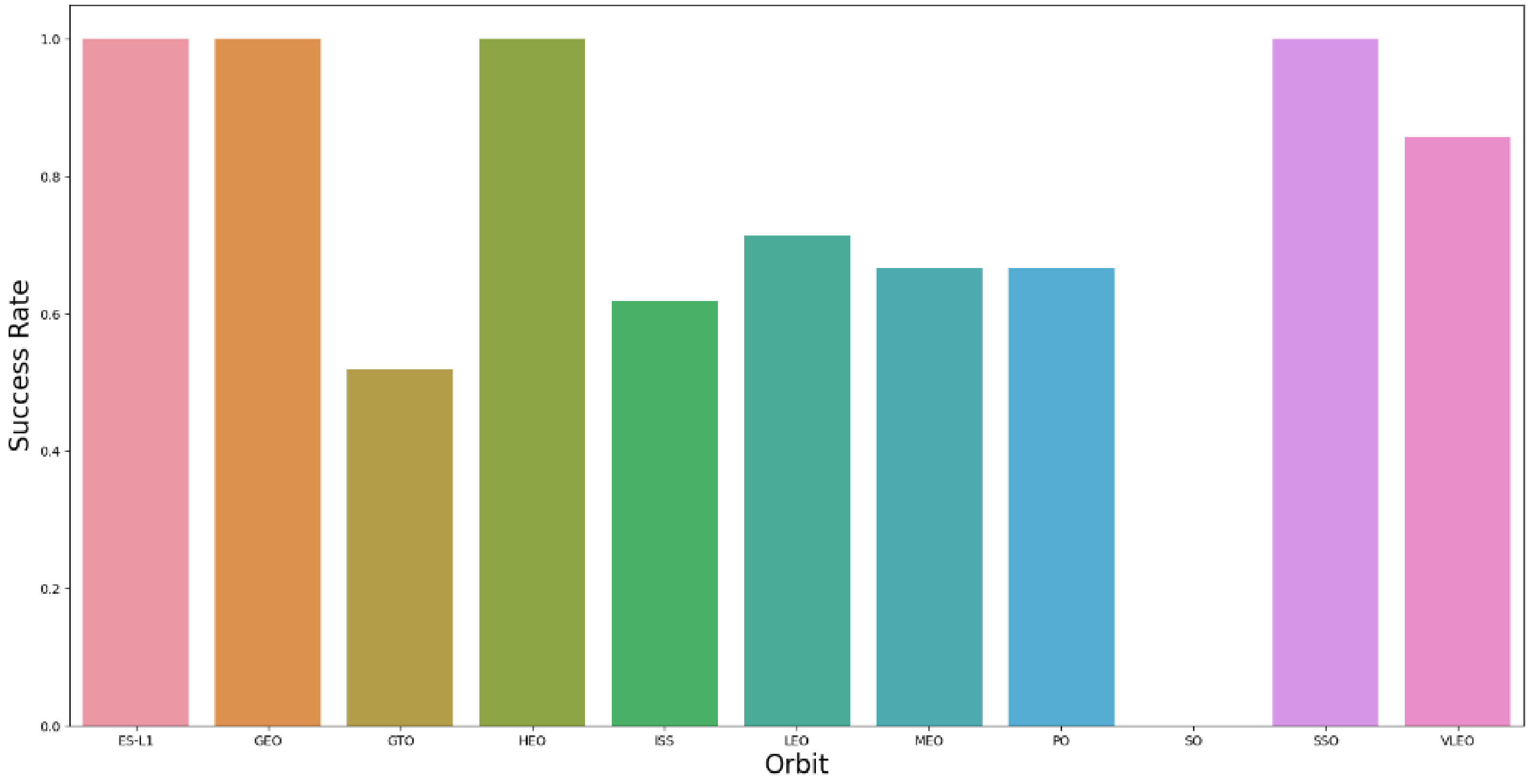
It seems that the more launches are made, the success rate increases.

Payload vs. Launch Site



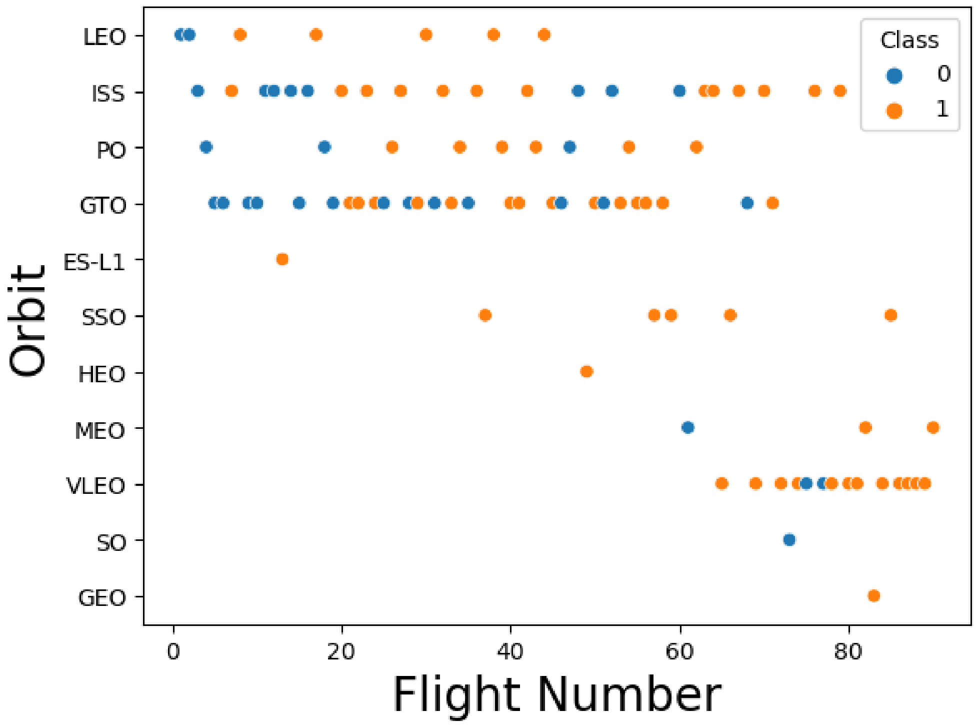
Now if you observe Payload Vs. Launch Site scatter point chart you will find for the VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000).

Success Rate vs. Orbit Type



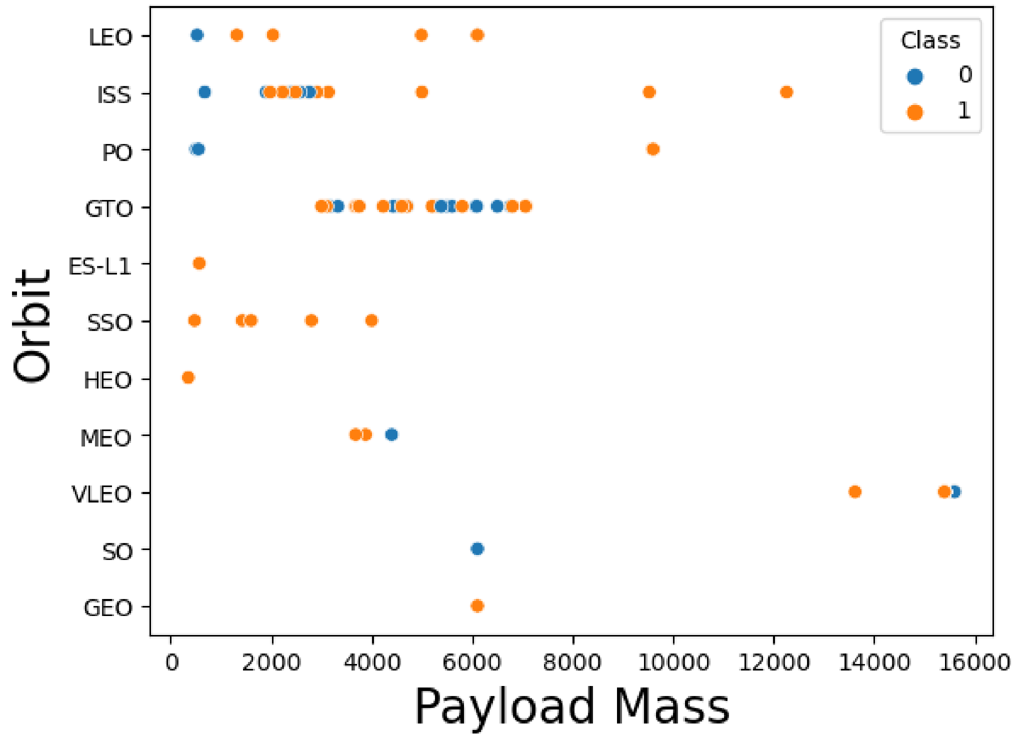
The ES-L1, GEO, HEO and SSO orbits have an excellent success rate, much higher than the rest.

Flight Number vs. Orbit Type



You should see that 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.

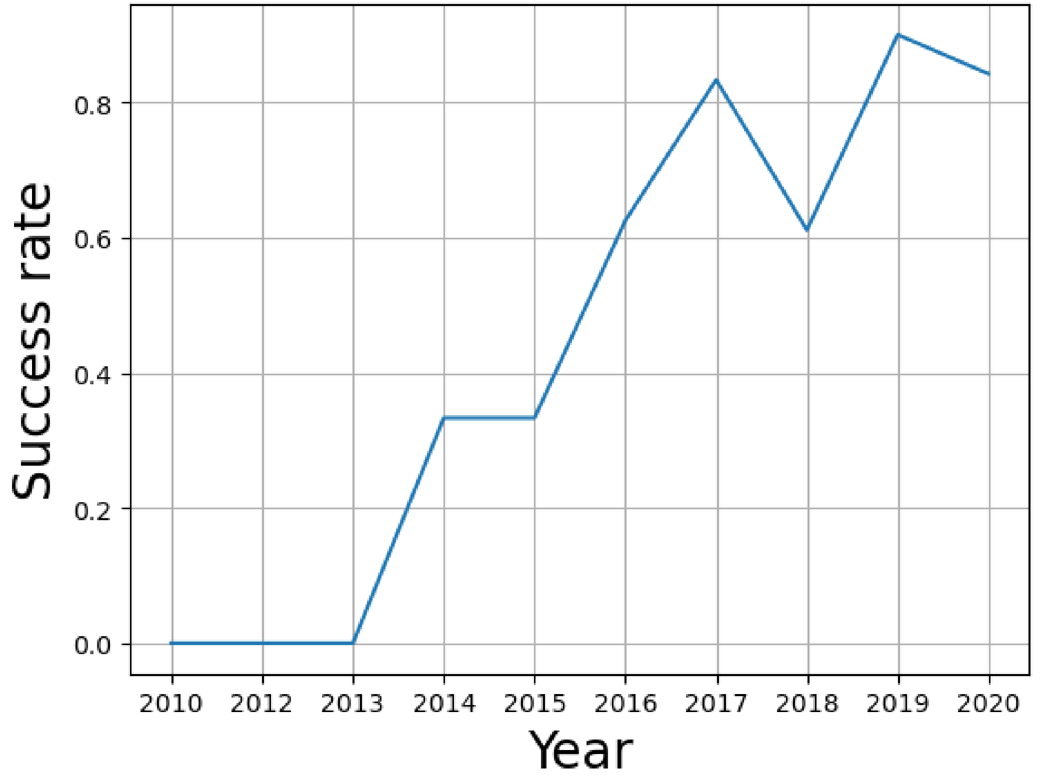
Payload vs. Orbit Type



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

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



You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.

All Launch Site Names

```
%%sql
SELECT DISTINCT Launch_site
FROM SPACEXTABLE;
```

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

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

Launch Site Names Begin with 'KSC'

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_site LIKE 'KSC%'
LIMIT 5;
```

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

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)
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-05-15	23:21:00	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

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

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS Total_Payload_Mass
FROM SPACEXTABLE
WHERE Customer LIKE '%NASA (CRS)%'
```

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

Total_Payload_Mass
48213

Average Payload Mass by F9 v1.1

```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS Average_Payload_Mass
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

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

Average_Payload_Mass
2928.4

First Successful Ground Landing Date

```
%%sql

SELECT *
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (drone ship)';
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
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)
2016-05-06	5:21:00	F9 FT B1022	CCAFS LC-40	JCSAT-14	4696	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
2016-05-27	21:39:00	F9 FT B1023.1	CCAFS LC-40	Thaicom 8	3100	GTO	Thaicom	Success	Success (drone ship)
2016-08-14	5:26:00	F9 FT B1026	CCAFS LC-40	JCSAT-16	4600	GTO	SKY Perfect JSAT Group	Success	Success (drone ship)
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)
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
2017-06-23	19:10:00	F9 FT B1029.2	KSC LC-39A	BulgariaSat-1	3669	GTO	Bulsatcom	Success	Success (drone ship)
2017-06-25	20:25:00	F9 FT B1036.1	VAFB SLC-4E	Iridium NEXT 2	9600	LEO	Iridium Communications	Success	Success (drone ship)
2017-08-24	18:51:00	F9 FT B1038.1	VAFB SLC-4E	Formosat-5	475	SSO	NSPO	Success	Success (drone ship)
2017-10-09	12:37:00	F9 B4 B1041.1	VAFB SLC-4E	Iridium NEXT 3	9600	Polar LEO	Iridium Communications	Success	Success (drone ship)
2017-10-11	22:53:00	F9 FT B1031.2	KSC LC-39A	SES-11 / EchoStar 105	5200	GTO	SES EchoStar	Success	Success (drone ship)
2017-10-30	19:34:00	F9 B4 B1042.1	KSC LC-39A	Koreasat 5A	3500	GTO	KT Corporation	Success	Success (drone ship)
2018-04-18	22:51:00	F9 B4 B1045.1	CCAFS SLC-40	Transiting Exoplanet Survey Satellite (TESS)	362	HEO	NASA (LSP)	Success	Success (drone ship)
2018-05-11	20:14:00	F9 B5 B1046.1	KSC LC-39A	Bangabandhu-1	3600	GTO	Thales-Alenia/BTRC	Success	Success (drone ship)

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Landing_Outcome = 'Success (ground pad)'
    AND PAYLOAD_MASS__KG_ > 4000
    AND PAYLOAD_MASS__KG_ < 6000;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
2017-09-07	14:00:00	F9 B4 B1040.1	KSC LC-39A	Boeing X-37B OTV-5	4990	LEO	U.S. Air Force	Success	Success (ground pad)
2018-01-08	1:00:00	F9 B4 B1043.1	CCAFS SLC-40	Zuma	5000	LEO	Northrop Grumman	Success (payload status unclear)	Success (ground pad)

Total Number of Successful and Failure Mission Outcomes

```
%%sql

-- Seleccionar el recuento de resultados exitosos y fallidos de las misiones donde los valores de Mission_Outcome comienzan con '
SELECT Mission_Outcome, COUNT(*) AS Total
FROM SPACEXTABLE
WHERE Mission_Outcome LIKE 'Success%' OR Mission_Outcome LIKE 'Failure%'
GROUP BY Mission_Outcome;
```

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

Mission_Outcome	Total
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE PAYLOAD_MASS__KG_ = (
    SELECT MAX(PAYLOAD_MASS__KG_)
    FROM SPACEXTABLE
);
```

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

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

2015 Launch Records

```
%%sql

SELECT
    substr(Date, 6, 2) AS Month,
    Mission_Outcome AS Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) = '2017'
    AND Landing_Outcome = 'Success (ground pad)';
```

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

Month	Landing_Outcome	Booster_Version	Launch_Site
02	Success	F9 FT B1031.1	KSC LC-39A
05	Success	F9 FT B1032.1	KSC LC-39A
06	Success	F9 FT B1035.1	KSC LC-39A
08	Success	F9 B4 B1039.1	KSC LC-39A
09	Success	F9 B4 B1040.1	KSC LC-39A
12	Success	F9 FT B1035.2	CCAFS SLC-40

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

```
%%sql

SELECT Landing_Outcome, COUNT(*) AS Outcome_Count
FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
AND Landing_Outcome IN ('Failure (drone ship)', 'Success (ground pad)')
GROUP BY Landing_Outcome
ORDER BY Outcome_Count DESC;
```

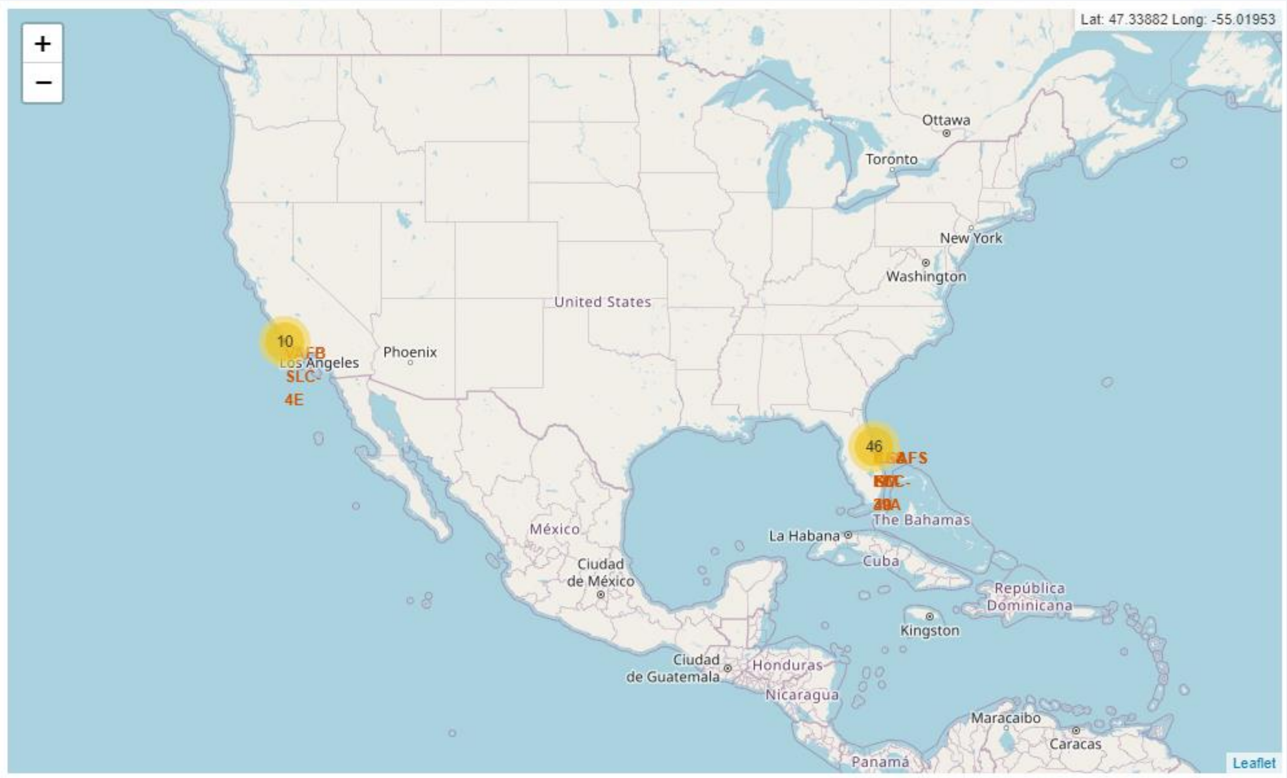
```
 * sqlite:///my_data1.db
Done.
```

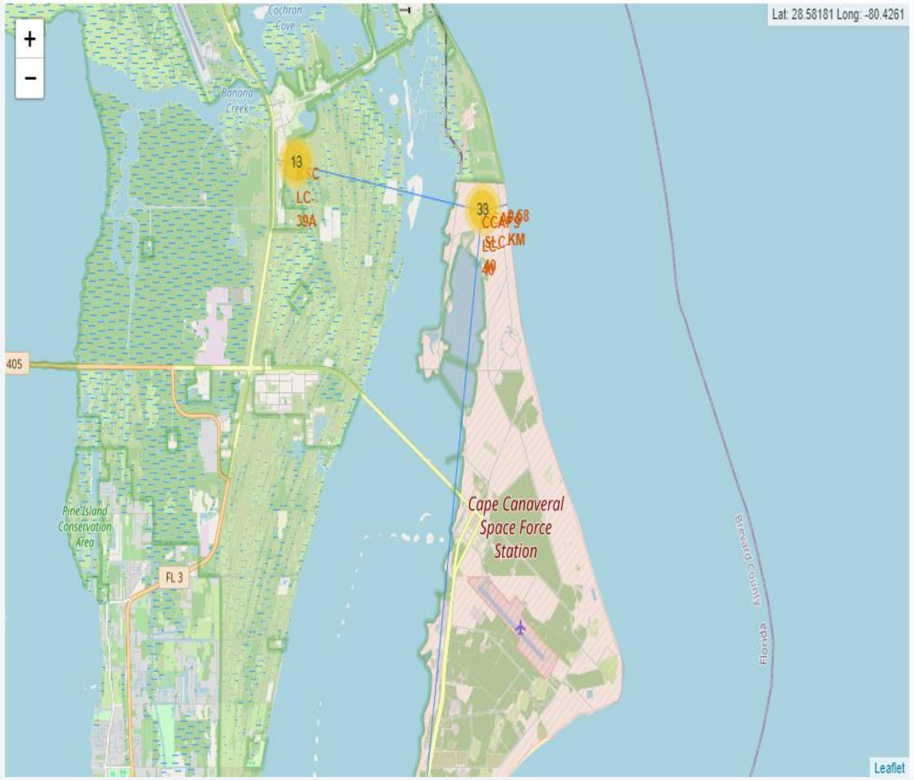
Landing_Outcome	Outcome_Count
Failure (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

MAIN LAUNCHSITES

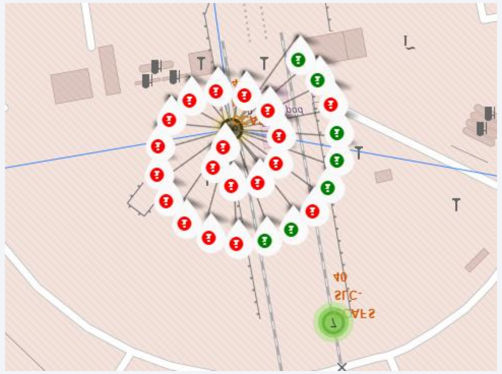




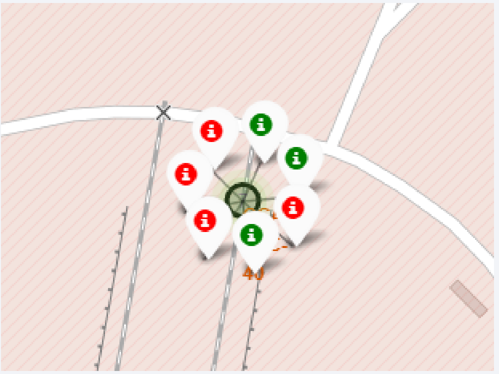
As we can see, the launch zones are always located near the ocean coast.

They are also found in areas where there is no urban density.

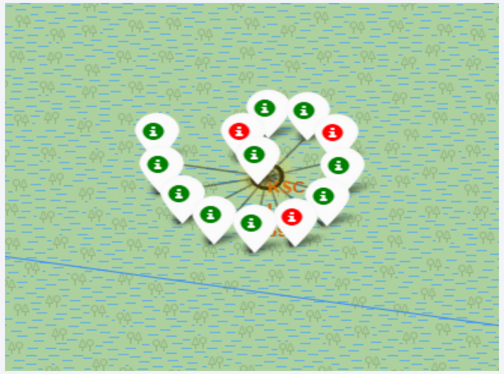
VISUAL SUCCESS RATE



CCAFS LC.40



CCAFS SLC 40



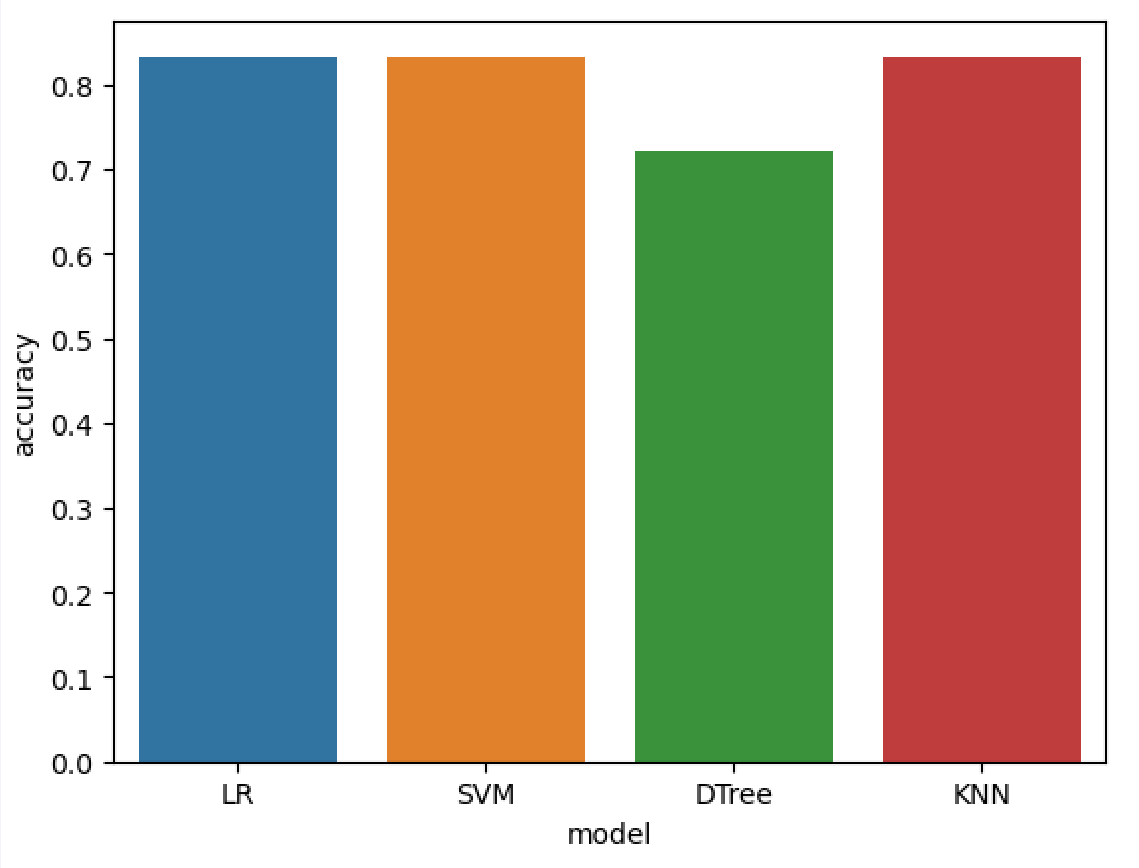
KSC LC-39A

VAFB SLC 4E

Section 5

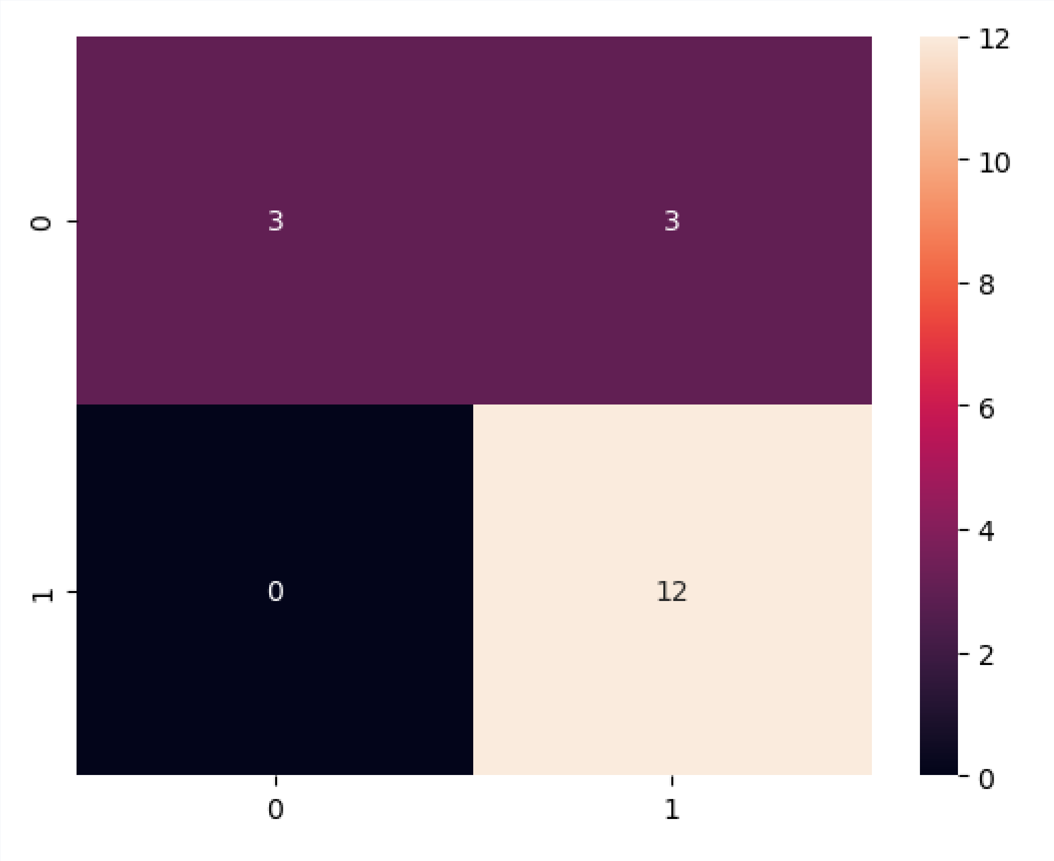
Predictive Analysis (Classification)

Classification Accuracy



- We have no difference in the performance of Log Reg, SVM and KNN. Decision Tree is discarded.

Confusion Matrix



- Examining the confusion matrix, we see that logistic regression can distinguish between the different classes. We see that the major problem is false positives.

Conclusions

- We have been able to see that both Payload Mass and Orbit are fundamental elements in the success of the launches
- ES-L1, GEO , HEO and SSO orbits have a 100% of success rate.
- Thanks to the interactive map visualization, we have been able to see that some launch points have a better success rate.
- KSC LC-39A is the launchsite with the best success rate.
- It seems tht the more launches are made, the success rate increases.
- Through machine learning we have been able to create a model which can predict with more than 80% accuracy whether the next launch will be successful or not.

Thank you!