Data Science Capstone Project

-Adityan Chanduruthil

Hello! I'm...

A human with a passion for turning caffeine into code.

.

Table of Contents.

Table of Contents v2.

1

Executive Summary

Summary of Methodologies

The Following Methodologies were used to analyse data:

- Data Collection using web scraping and SpaceX API.
- Exploratory Data Analysis (EDA) including data wrangling , data visualisation and interactive visual analytics.
- Machine Learning Prediction (Classification)

Summary of All Results

- It was possible to Collect valuable data from public source.
- EDA allowed to identify which features are best to predict success of launchings.
- Machine Learning Prediction showed the best model to predict which characteristics are important to drive this opportunity by the best way , using all collected data.

2

Introduction

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 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.

Questions to be Answered

How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?

Does the rate of successful landings increase over the years?

What is the best algorithm that can be used for binary classification in this case?

3

Methodology

Methodology

Data collection methodology:

- Using SpaceX Rest API

- Using Web Scraping from Wikipedia

Performed data wrangling:

- Filtering the data

- Dealing with missing values

- Using One Hot Encoding to prepare the data to a binary classification

Methodology

Performed exploratory data analysis (EDA) using visualization and SQL

Performed interactive visual analytics using Folium and Plotly Dash

Performed predictive analysis using classification models

- Building, tuning and evaluation of classification models to ensure the best

Data Collection

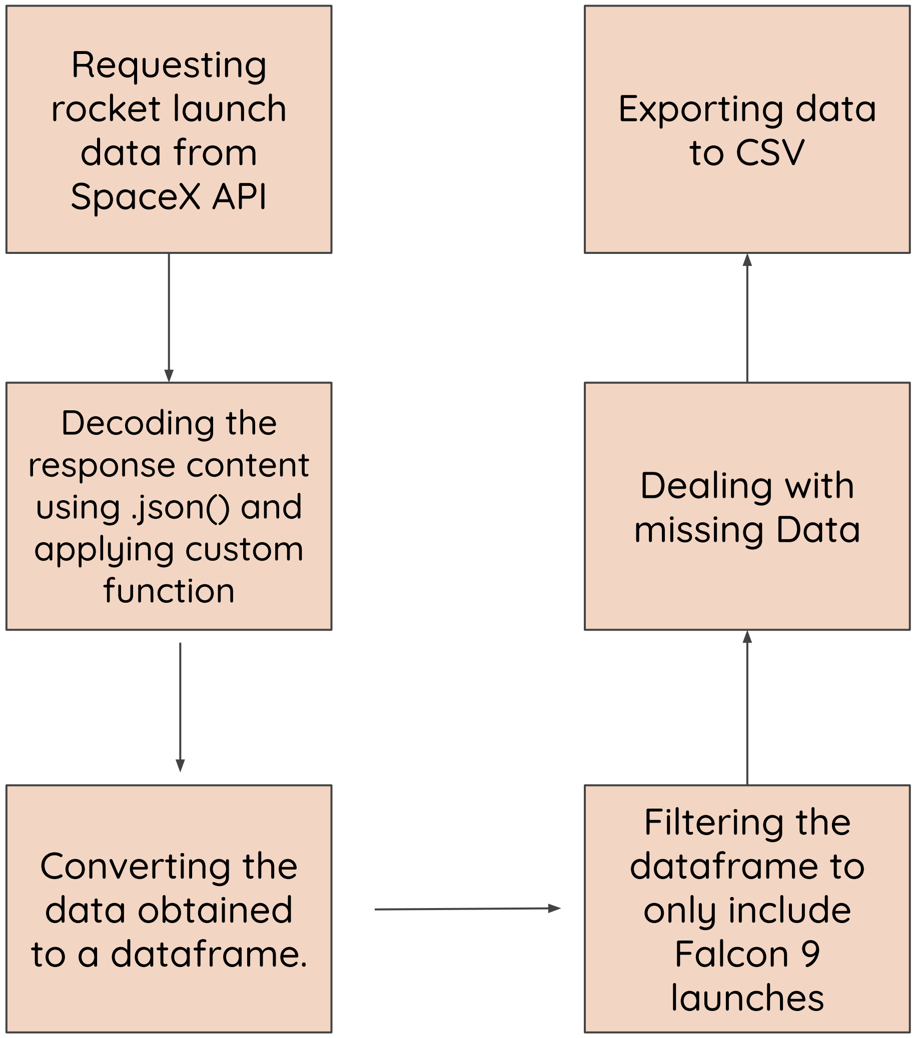
Data collection process involved a combination of API requests from SpaceX REST API and Web Scraping data from a table in SpaceX’s Wikipedia entry.

We had to use both of these data collection methods in order to get complete information about the launches for a more detailed analysis.

SpaceX API

SpaceX offers a public API from where data can be obtained and used.

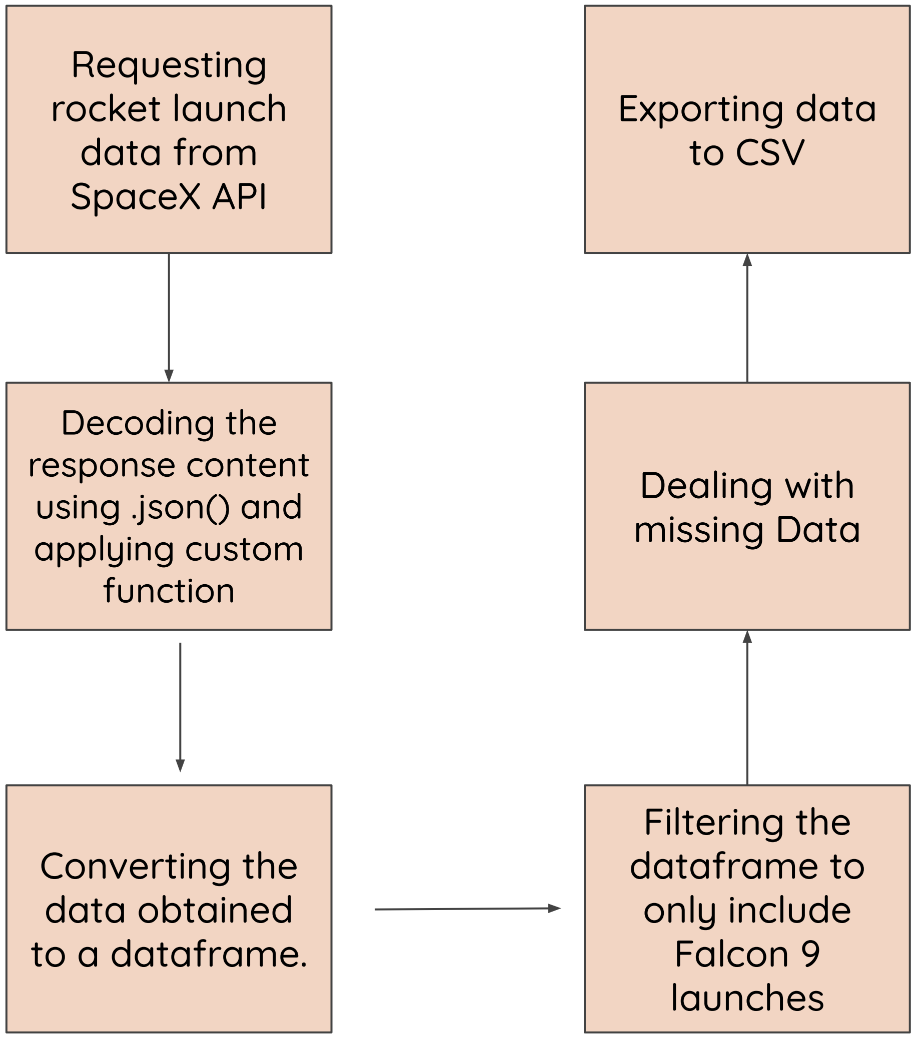
This API was used according to the flowchart beside and then data is persisted.



Web Scraping

Data from SpaceX launches can also be obtained from wikipedia.

Data are downloaded from Wikipedia according to the flowchart and then persisted.

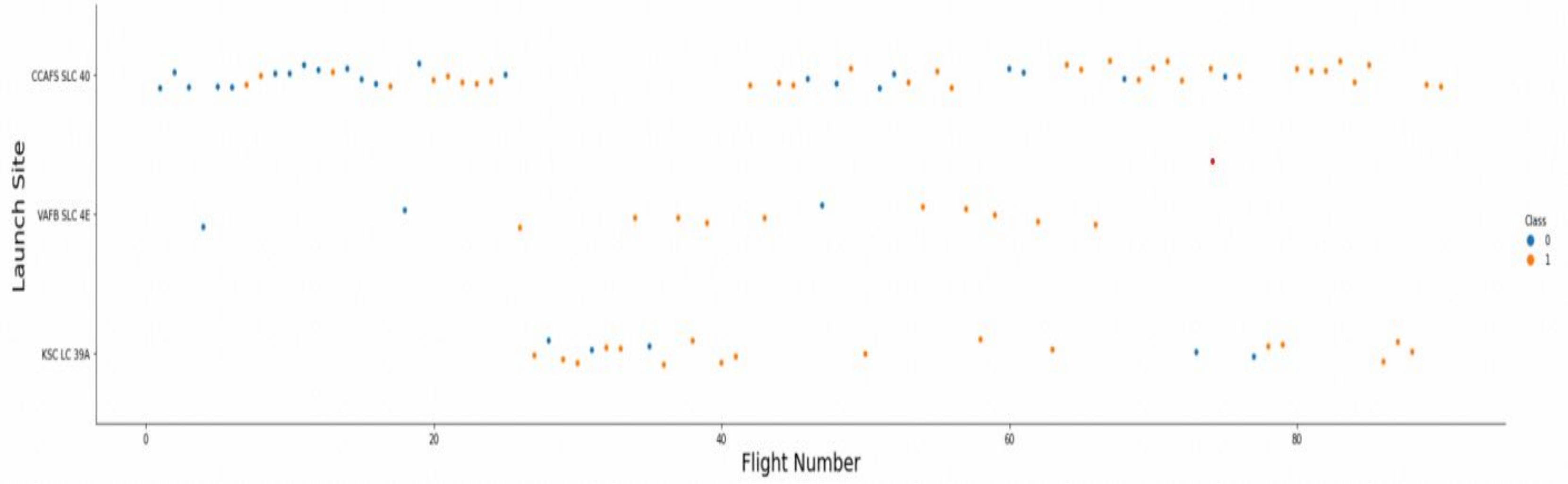


Data Wrangling

- Initially some Exploratory Data Analysis (EDA) was performed on the dataset .
- Then the summaries launches per site, occurrences of each orbit and occurrences of mission outcome per orbit type were calculated.
- Finally , the landing outcome label was created from outcome column

EDA with Visualisation

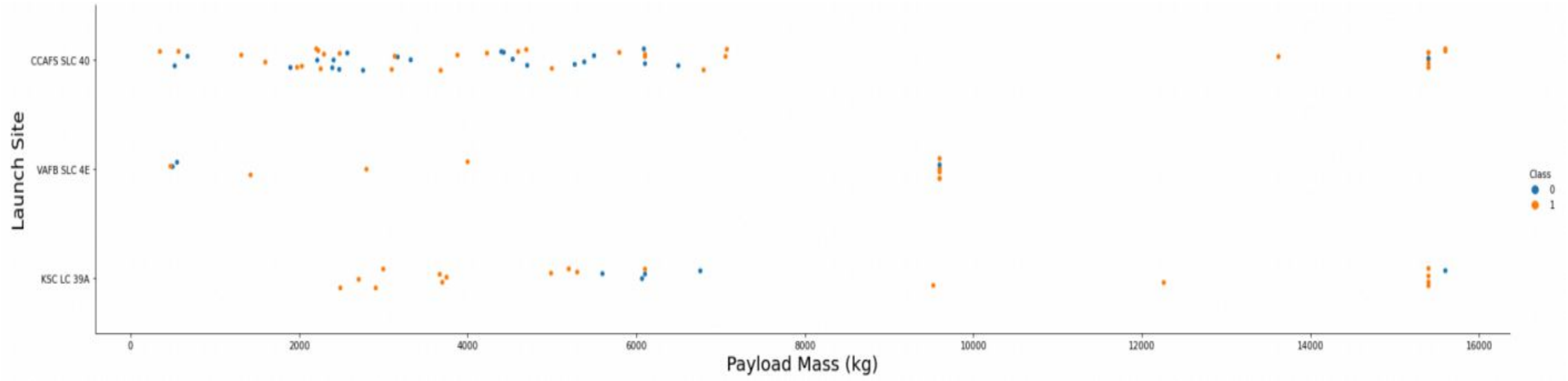
Flight Number vs Launch Site



Observations:

- The earliest flights all failed while the latest flights all succeeded.
- The CCAFS SLC 40 launch site has about a half of all launches.
- VAFB SLC 4E and KSC LC 39A have higher success rates.
- It can be assumed that each new launch has a higher rate of success.

Payload vs Launch Site



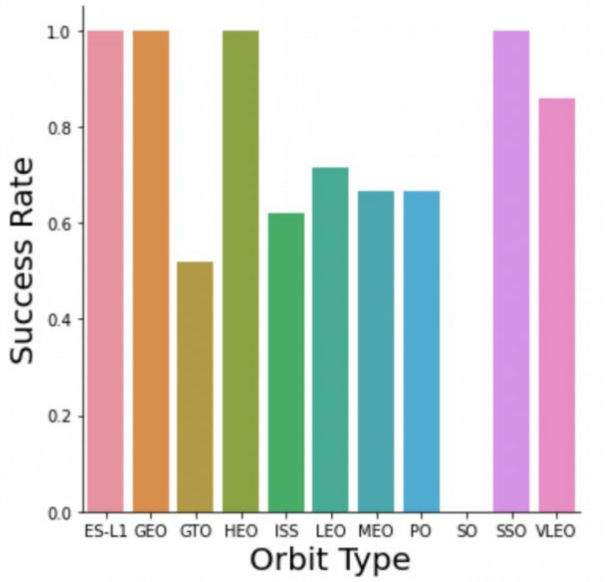
Observations:

- For every launch site the higher the payload mass, the higher the success rate.
- Most of the launches with payload mass over 7000 kg were successful.
- KSC LC 39A has a 100% success rate for payload mass under 5500 kg too.

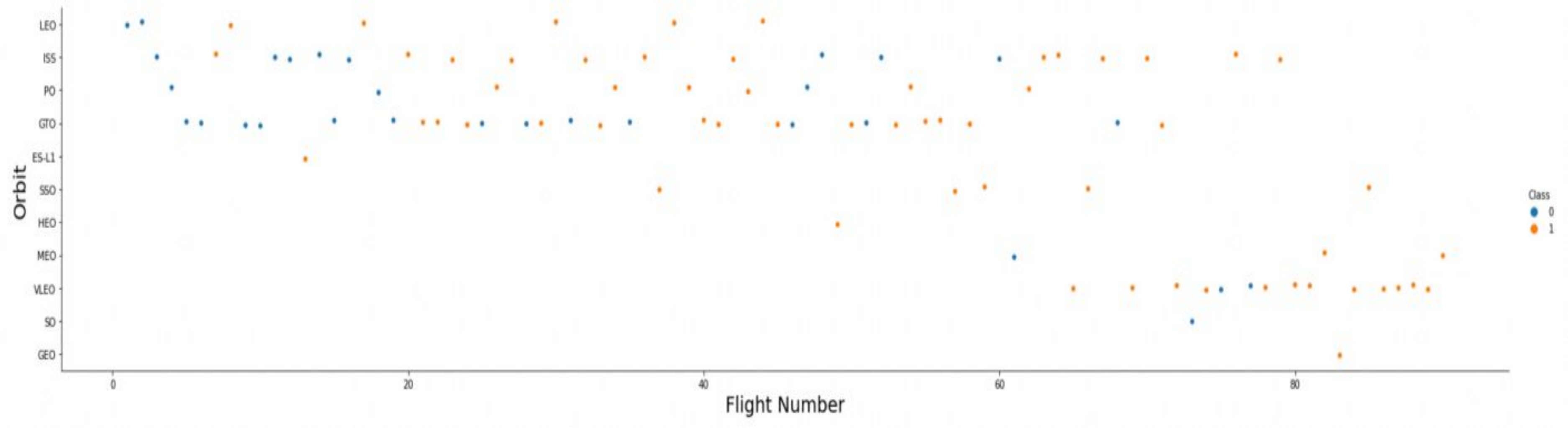
Success rate vs Orbit type

Observations:

- Orbits with 100% success rate: ES-L1, GEO, HEO, SSO
- Orbits with 0% success rate: SO
- Orbits with success rate between 50% and 85%: GTO, ISS, LEO, MEO, PO



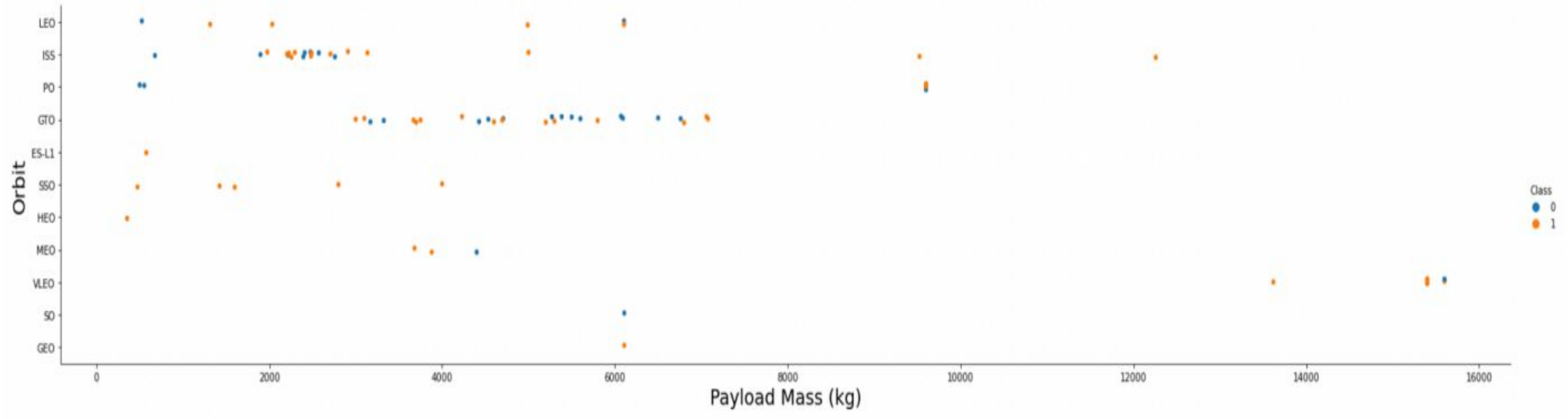
Flight NUmber vs Orbit Type



Observations:

- 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.

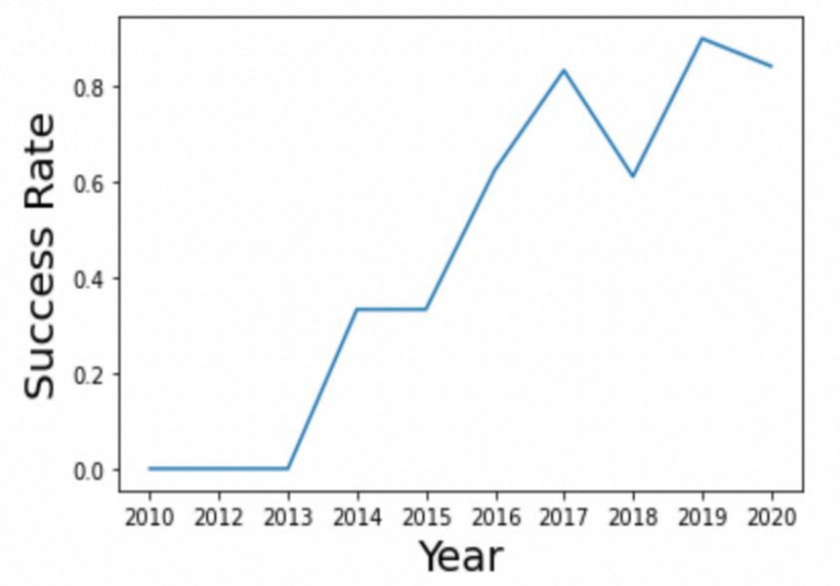
Payload Mass vs Orbit Type



Observations:

- Heavy payloads have a negative influence on GTO orbits and positive on GTO and Polar LEO (ISS) orbits.

Launch Success yearly Trend



Observations:

- Success rate since 2013 kept increasing till 2020

EDA with SQL

All Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[4]:
```

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

Observations:

- Displaying the names of the unique launch sites in the space mission.

Launch Site Names begin with "CCA"

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

Observations:

- Displaying 5 records where launch sites begin with the string 'CCA'.

Total Payload Mass

In [6]:
```
%sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

Observations:

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

Average payload mass by F9v1.1

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[7]:
```

average_payload_mass
2534

Observations:

- Displaying average payload mass carried by booster version F9 v1.1.

First Successful ground landing date

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[8]:
```

first_successful_landing
2015-12-22

Observations:

- Listing the date when the first successful landing outcome in ground pad was achieved .

Successful drone ship landing with Payload b/w 4000 and 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;

         * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
        Done.

Out[9]:
```

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

Observations:

- Listing the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Total Number of successful and failure mission outcomes

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[10]:
```

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Observations:

- Listing the total number of successful and failure mission outcomes.

Boosters carried maximum payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[11]:
```

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

Observations:

- Listing the names of the booster versions which have carried the maximum payload mass

2015 Launch Records

In [12]:
```
%%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
      where landing__outcome = 'Failure (drone ship)' and year(date)=2015;
```

```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Observations:

- Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

Rank success count between 2010-06–04 and 2017-03-20

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
               where date between '2010-06-04' and '2017-03-20'
               group by landing__outcome
               order by count_outcomes desc;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
Out[13]:
```

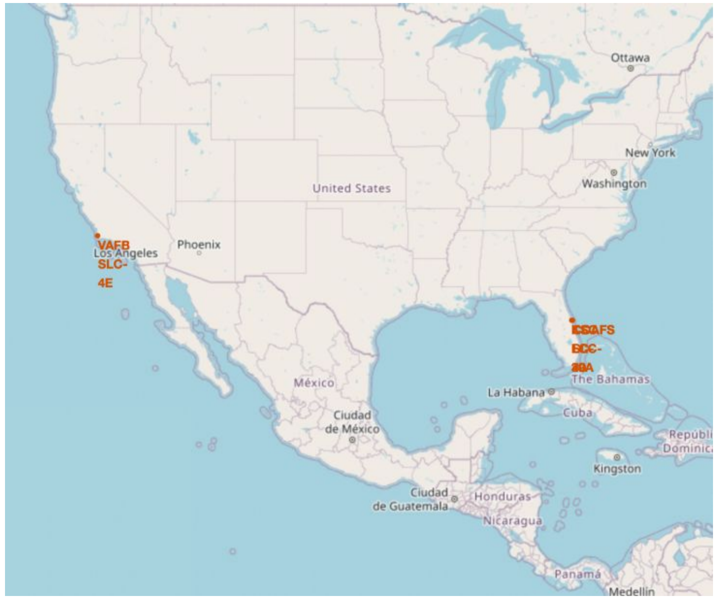
landing__outcome	count_outcomes
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Observations:

- Ranking 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.

Interactive Map with Folium

All launch sites' location markers on a global map



Observations/Explanation :

Launch sites near the equator provide a speed boost due to the Earth's faster rotation.Objects at the equator already move at 1670 km/hr, helping spacecraft maintain orbital speed. Proximity to the coast minimizes risks by directing rocket launches toward the ocean, reducing the chance of derbis affecting populated areas.

Colour-labeled launch records on the map

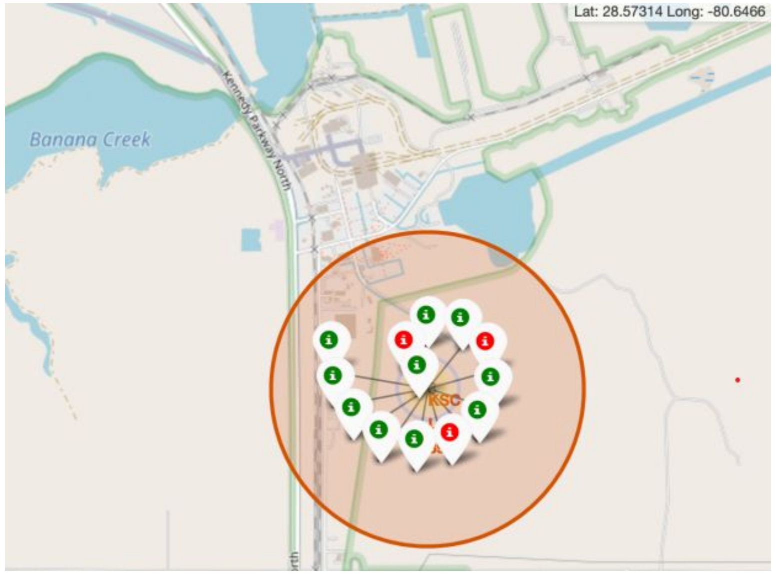
Observations/Explanation :

From the colour-labeled markers we should be able to easily identify which launch sites have relatively high success rates.

- Green Marker = **Successful** Launch

- Red Marker = **Failed** Launch

• Launch Site **KSC LC-39A** has a very high Success Rate.



Distance from the launch site KSC LC-39A to its proximities

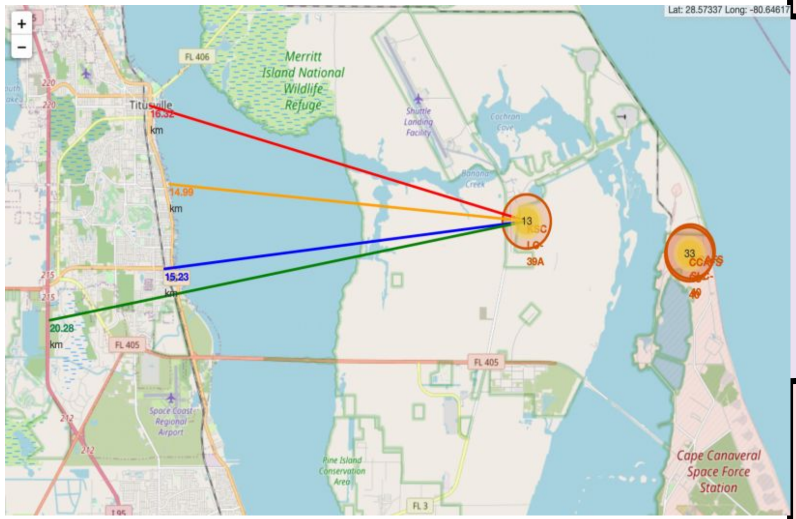
Observations/Explanation :

From the visual analysis of the launch site **KSC LC-39A** we can clearly see that it is:

• relative close to railway (15.23 km) - relative close to highway (20.28 km) relative close to coastline (14.99 km)

• Also the launch site KSC LC-39A is relative close to its closest city Titusville (16.32 km).

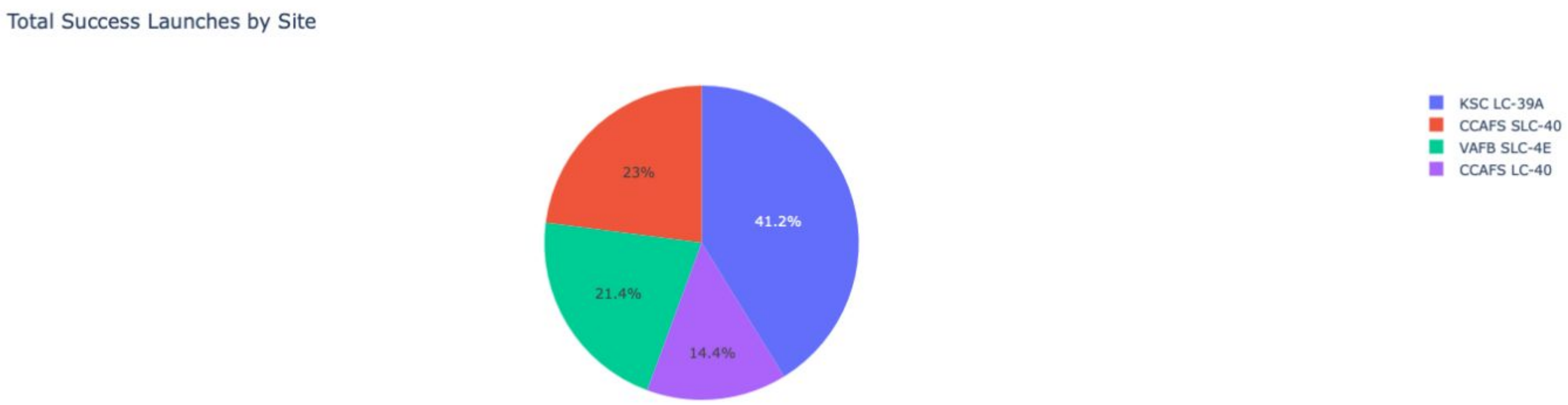
• Failed rocket with its high speed can cover distances like 15-20 km in few seconds. It could be potentially dangerous to populated areas.



.

Build a DashBoard with Plotly Dash

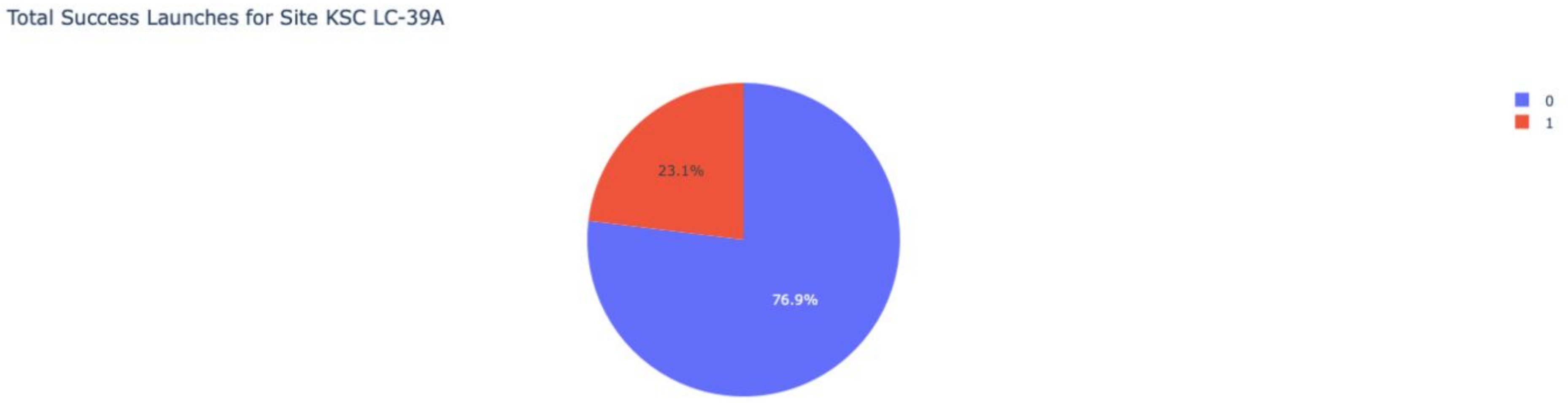
Launch success count for all sites



Observations/Explanation :

- The chart clearly shows that from all the sites, KSC LC-39A has the most successful launches

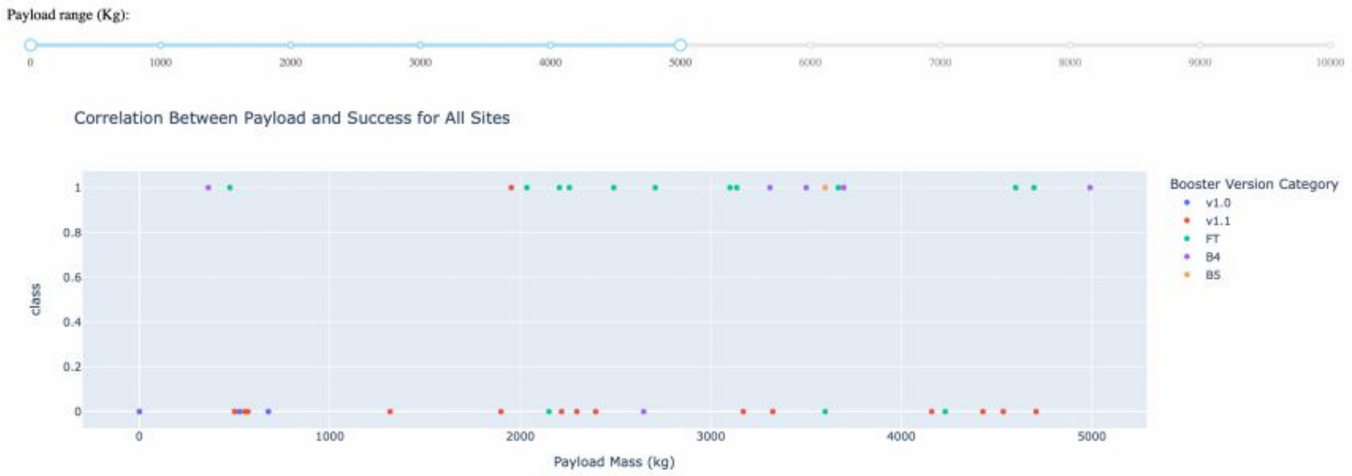
Launch site with highest launch success ratio

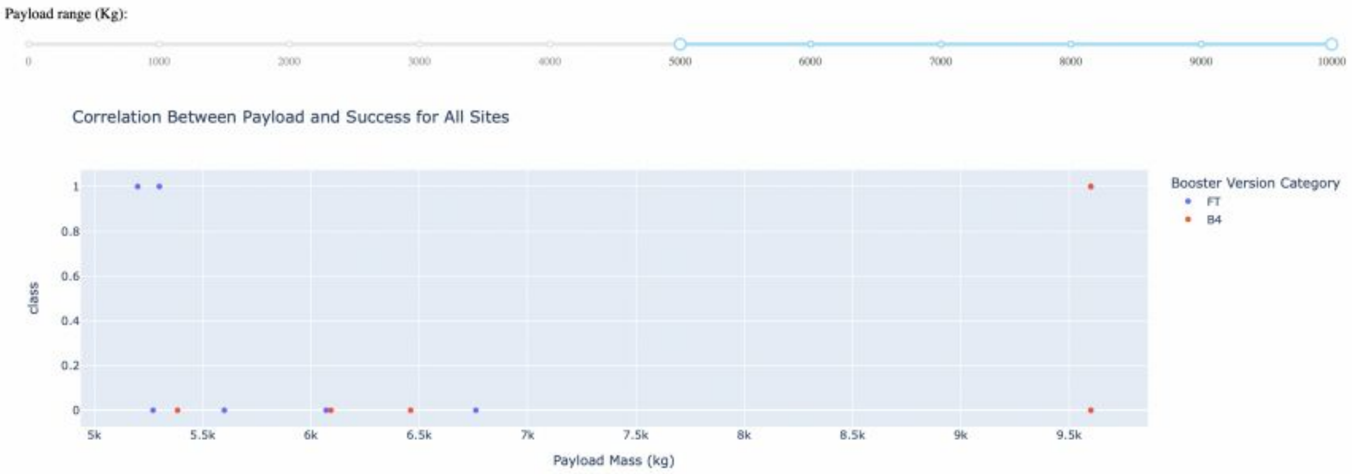


Observations/Explanation :

- KSC LC-39A has the highest launch success rate (76.9%) with 10 successful and only 3 failed landing.

Payload Mass vs. Launch Outcome for all sites





Observations/Explanation :

- The charts show that payloads between 2000 and 5500 kg have the highest success rate

.

Predictive Analysis (Classification)

Classification Accuracy

Observations/Explanation :

•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. This model has not only higher scores, but also the highest accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

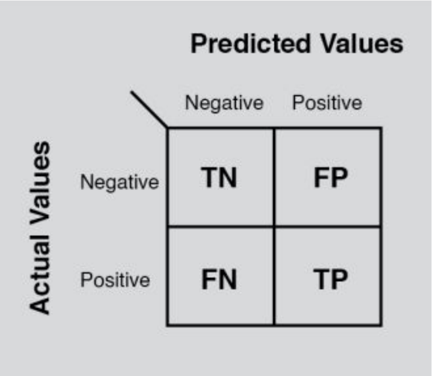
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

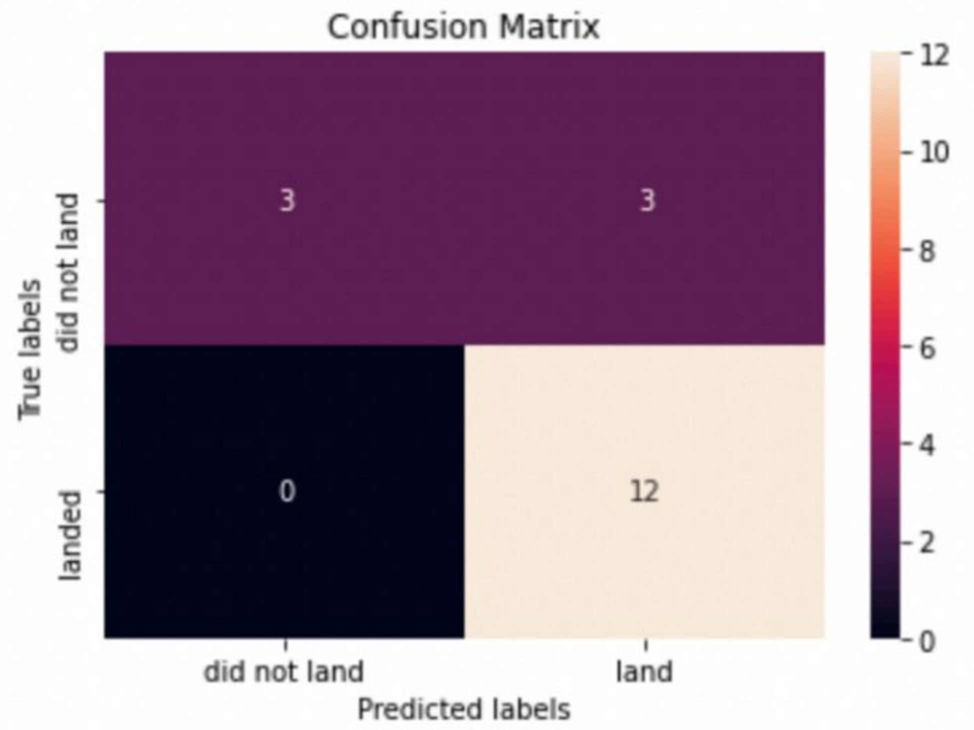
Confusion Matrix

Observations/Explanation :

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

.





5

Conclusion

Conclusions

Decision Tree Model Dominance

The Decision Tree model emerges as the most effective algorithm for predicting launch success within this dataset, showcasing its versatility and robust performance

Payload Mass Impact

Interestingly, launches with lower payload masses demonstrate higher success rates compared to their heavier counterparts. This suggests that managing and optimizing payload mass could play a crucial role in ensuring mission success.

Geographical Considerations

The strategic placement of launch sites near the Equator line and coastal regions is evident. Proximity to the Equator provides a natural boost for launching, and coastal locations facilitate water-based landing options

Success Rate Evolution Over Years

A positive trend is observed in the success rate of launches over the years. This could be indicative of advancements in technology, improved operational procedures, and a growing understanding of potential challenges.

KSC LC-39A Excellence

Among the launch sites, KSC LC-39A stands out with the highest success rate. Investigating the factors contributing to this success could provide valuable insights for other launch facilities.

Orbit-Specific Success

Orbits such as ES-L1, GEO, HEO, and SSO exhibit a remarkable 100% success rate. Understanding the specific characteristics of these orbits and applying similar strategies to other missions may contribute to overall mission success.

In summary, leveraging the strengths of the Decision Tree model, optimizing payload masses, and learning from the success factors of specific launch sites and orbits could further enhance the overall success rate of space missions.

6

Appendix

Credits.

Special Thanks to:

Ibm Instructors

Coursera