Falcon 9 - First Stage Landing Prediction

Héctor Alarcón Martínez
21.01.2025

Outline

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

Introduction

SpaceX is an American space technology company that has made remarkable advances in rocket propulsion, satellite communications, human spaceflight, and reusable launch vehicles, becoming the leading provider of space transportation thanks to its relatively low-cost launches.

A pivotal key to this cost reduction is the ability to land the first stage of the rocket, the booster, which accounts for about 70% of the total cost of the launch.

In this presentation, wiki data extracted through web scraping and the SpaceX API will be analyzed to obtain insights and attempt to predict the safe landing of the boosters using a machine learning model trained on this public data.

Executive Summary

- **Summary of methodologies used to analyze data:**

  Data collection, and wrangling, through SpaceX REST API and web scraping.

  Exploratory Data Analysis (EDA) with SQL.

  Data visualization using a dashboard made with Dash and Plotly and an interactive map with Folium.

  Several Machine Learning (ML) prediction techniques to find which features determine a successful launch.

- **Summary of all results:**

  The best features to predict launch success have been identified.

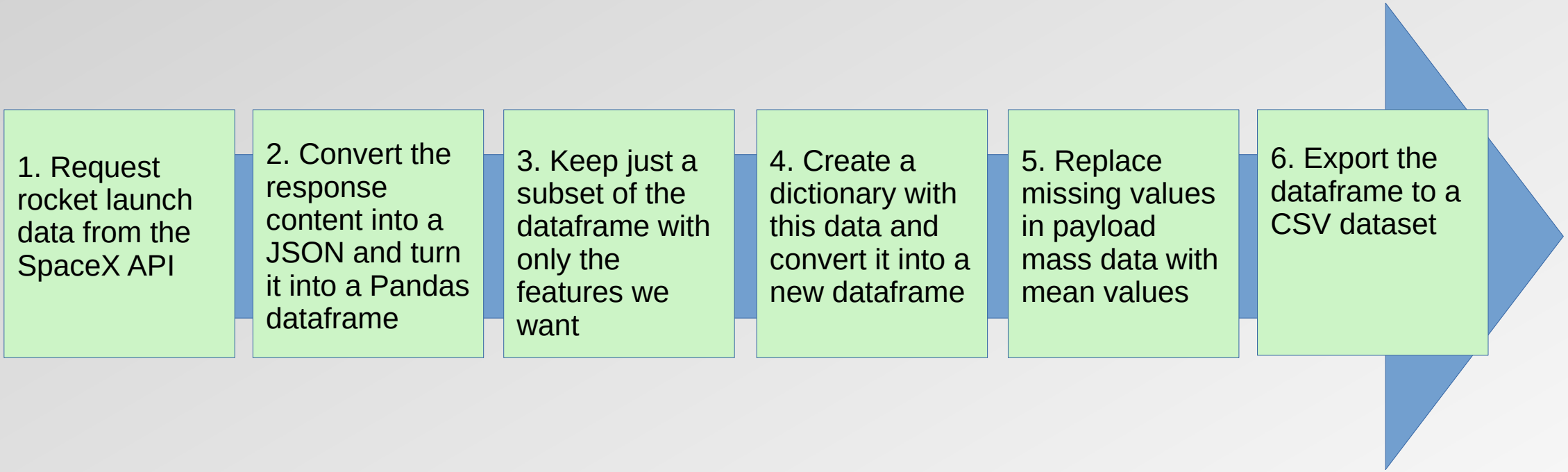
  ML analysis showed the best model for predicting optimal features for successful launches.

Methodology

Methodology

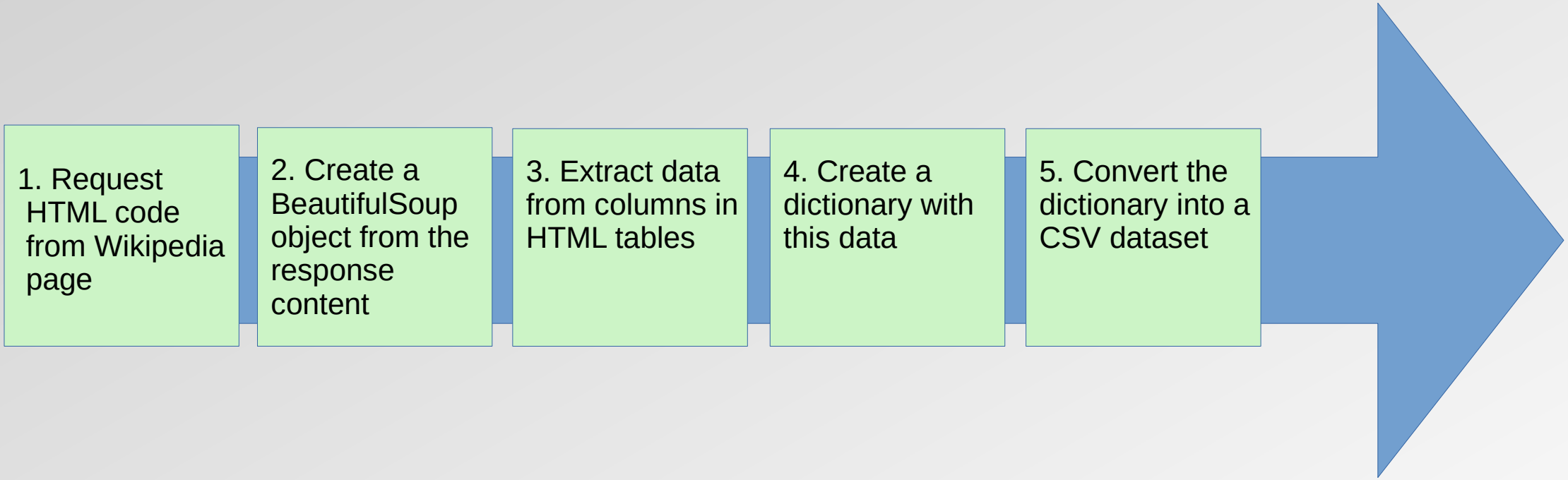
- Data Collection
  - All data was extracted from the SpaceX REST API and Wikipedia.
- Data Wrangling
  - The data was converted to several CSV tables.
  - A training label showing if the booster successfully landed was made.
- Exploratory Data Analysis (EDA) using visualization and SQL
- Interactive Visual Analytics using a Map and a Dashboard
- Predictive Analysis using several Machine Learning Models

Data Collection: SpaceX API



Python Notebook with code and results.

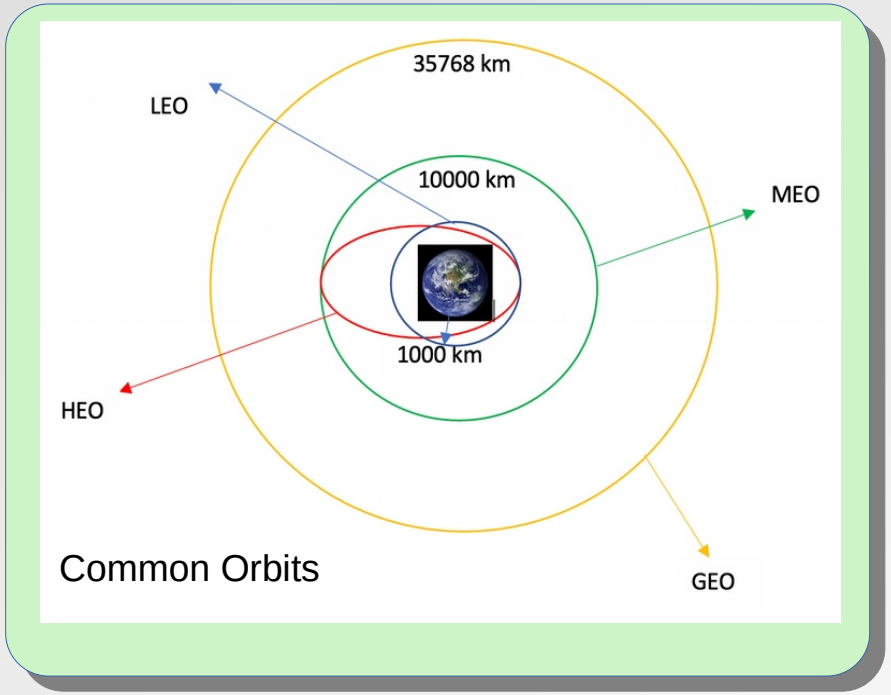
Data Collection: Web Scraping



Python Notebook with code and results.

Data Wrangling

There are several cases where the rocket failed to land successfully. Sometimes it manages to return but fails in the landing attempt due to an accident. Each launch is also associated with a certain type of orbit. In the dataset, these cases are reflected by a combination of several variables that, in order to subsequently train an ML model that predicts the success or failure of a mission, are converted to a new training label "Class" where 1 means that the booster landed successfully and 0 the opposite.



Python Notebook with code and results.

EDA with Data Visualization

Several scatter plots, bar plots and line charts were plotted in order to determine the correlations between many combinations of variables.

Scatter plots:

Flight Number vs, Launch Site

Payload Mass vs. Launch Site

Flight Number vs. Orbit Type

Payload Mass vs. Orbit Type

Bar Graph:

Success Rate vs. Orbit Type

Line Graph:

Success Rate vs. Year

Python Notebook with code and results.

EDA with SQL

Multiple SQL queries were executed to gather insights about the dataset such as:

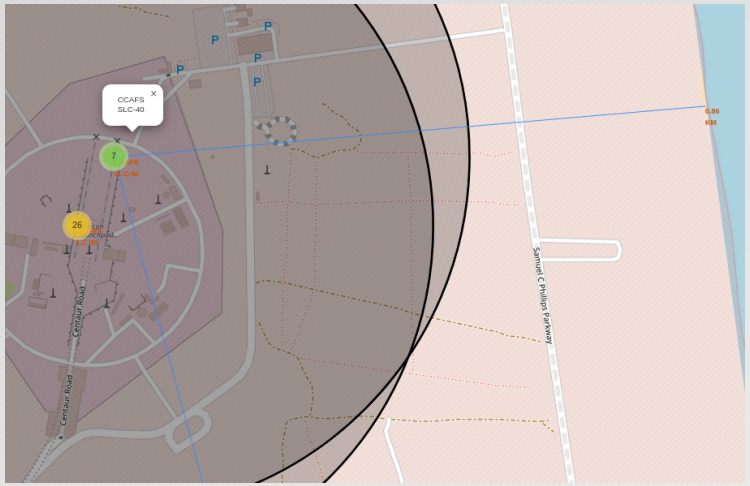
- **Find all launch site names.**
- **Find launches from launch sites whose name begin with ‘CCA’.**
- **Find the total payload mass carried by boosters launched for NASA missions.**
- **Find the average payload mass carried by the booster F9 v1.1**
- **Find when the first successful ground landing was achieved.**
- **List the boosters which have succeeded in drone ship landing carrying payloads between 4000 kg and 6000 kg.**
- **Find the total number of successful and failure mission outcomes.**
- **List the booster versions which have carried the maximum payload mass.**
- **Find the failed landing outcomes of the year 2015, indicating the month, launching site and booster rocket version.**
- **Rank the counts of landing outcomes between 2010-06-04, and 2017-03-20, in descending order by quantity.**

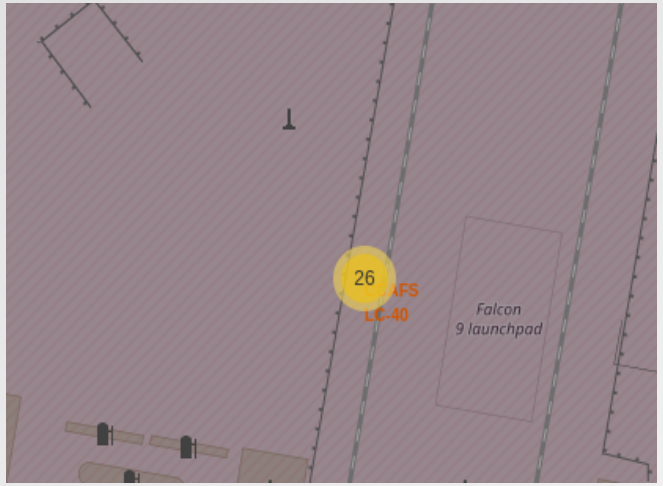
Python Notebook with code and results.

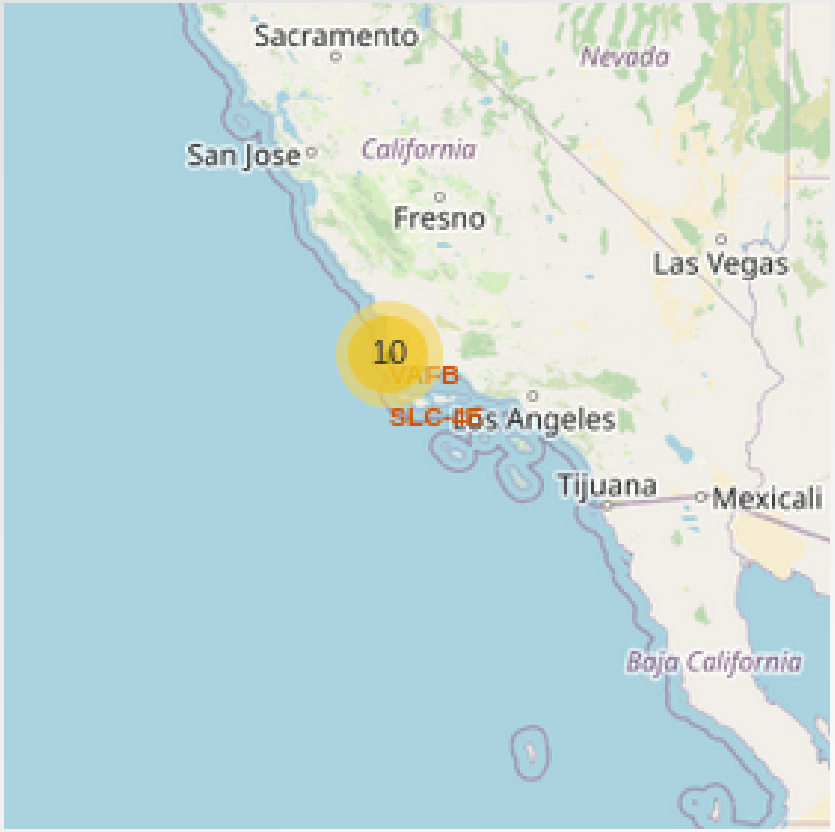
Interactive Visual Analytics – Launch Sites Map

An interactive geographic visualization was implemented, using Python's Folium library, showing the locations of all launch sites along with the number of launches and whether or not they were successful.

In addition, some distances to important landmarks (such as railways or coastlines) were calculated to provide insight into possible reasons for the general locations of these launches.







Python Notebook with code and results.

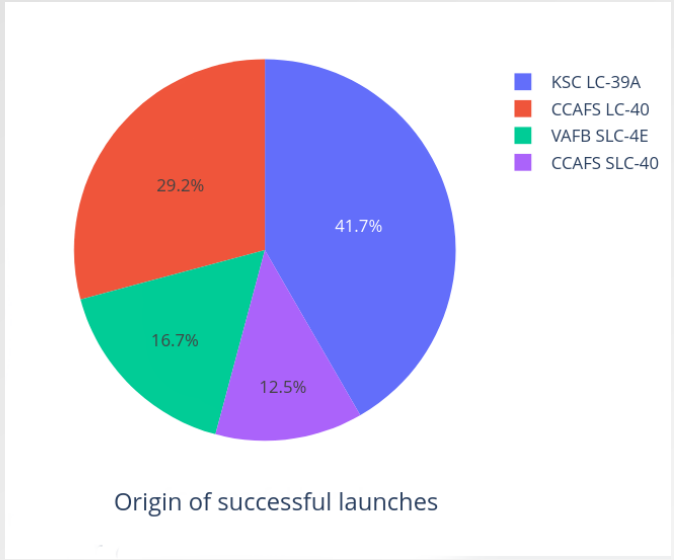
Interactive Visual Analytics – Dashboard

A website was implemented showing an interactive Dashboard, using Python and the Dash library, which interactively displays some graphs showing diverse statistics about the launches, their locations, quantity and whether they were successful or not.

A drop-down list of launch sites allows the user to filter data for each particular launch site or for all of them.

Pie charts display information about launches in general or from a particular selected site.

A scatter chart shows the correlation between payload mass and the outcome of each mission. Different payload ranges can be selected using a slider bar.



Python code at GitHub.

Predictive Analysis (Classification)

Using **scikit-learn**, four Machine Learning (ML) models were trained on the dataset (previously divided into a **training set** and a **test set**) in order to find which one model perform the best.

The hyperparameters were tuned by iterating several times until the classification accuracy of the models improved a little.

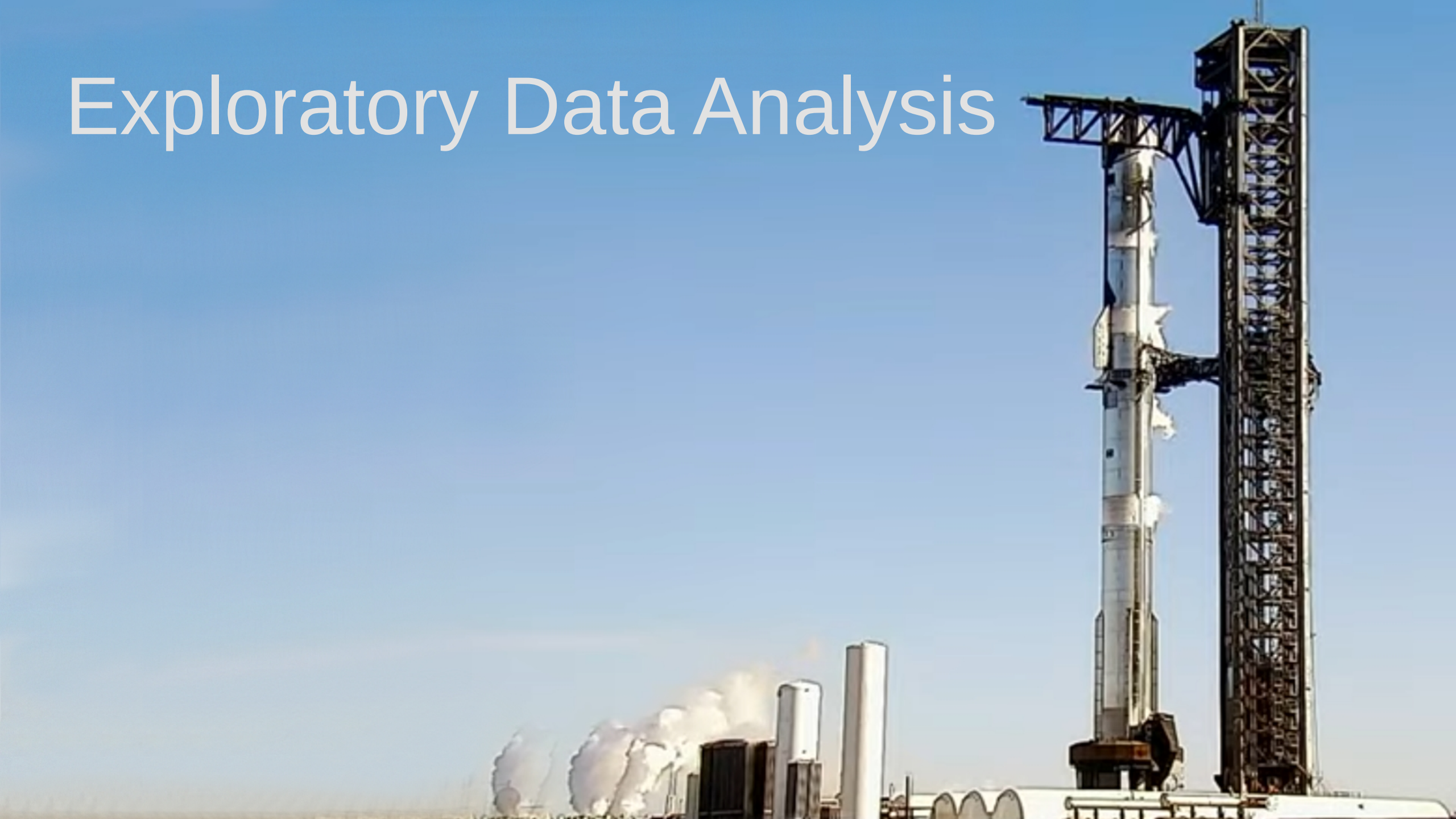
We then tested the models on the **test set** to find which one has the best performance.

Python Notebook with code and results.

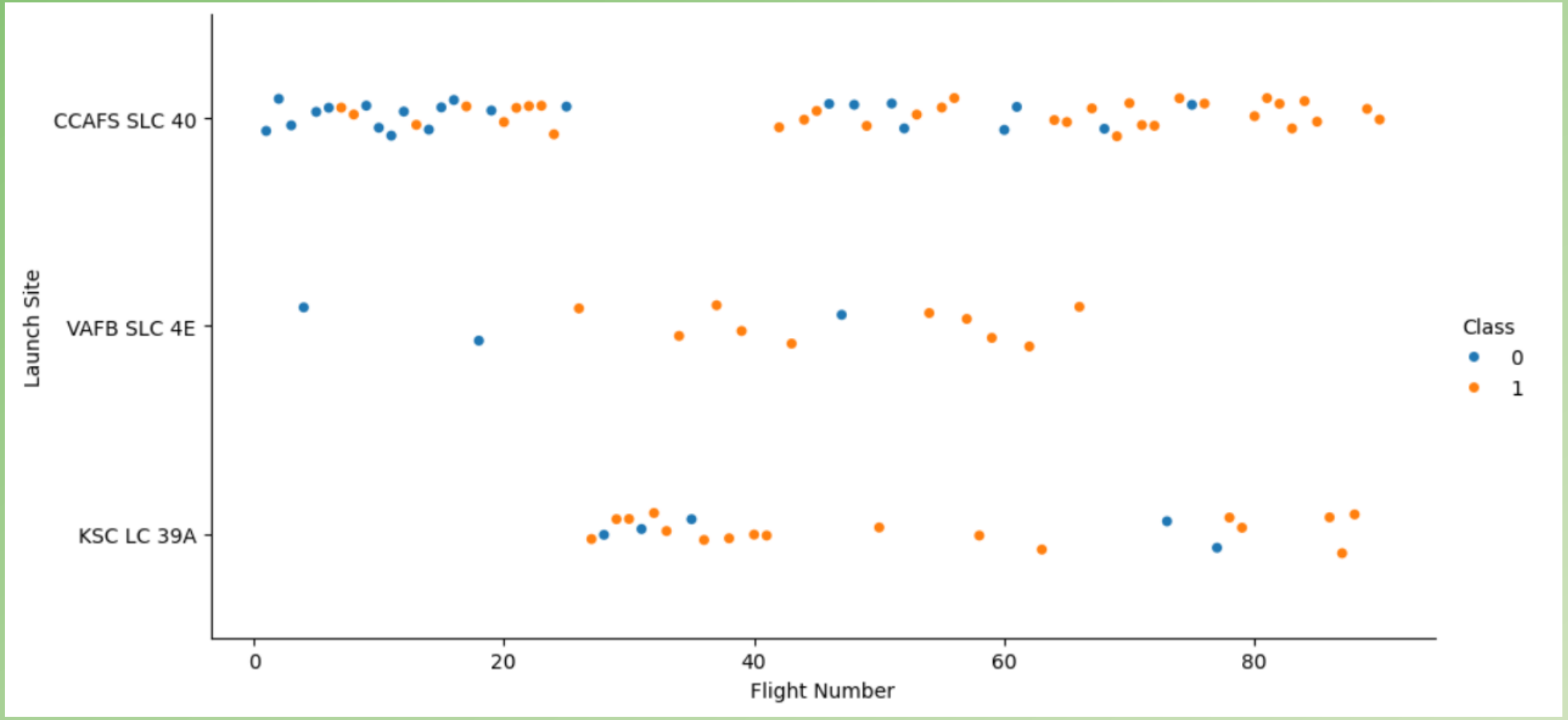
Results

- Exploratory Data Analysis
- Interactive Analysis: Proximity Map
- Interactive Analysis: Dashboard
- Predictive Analysis
- Conclusions

Exploratory Data Analysis

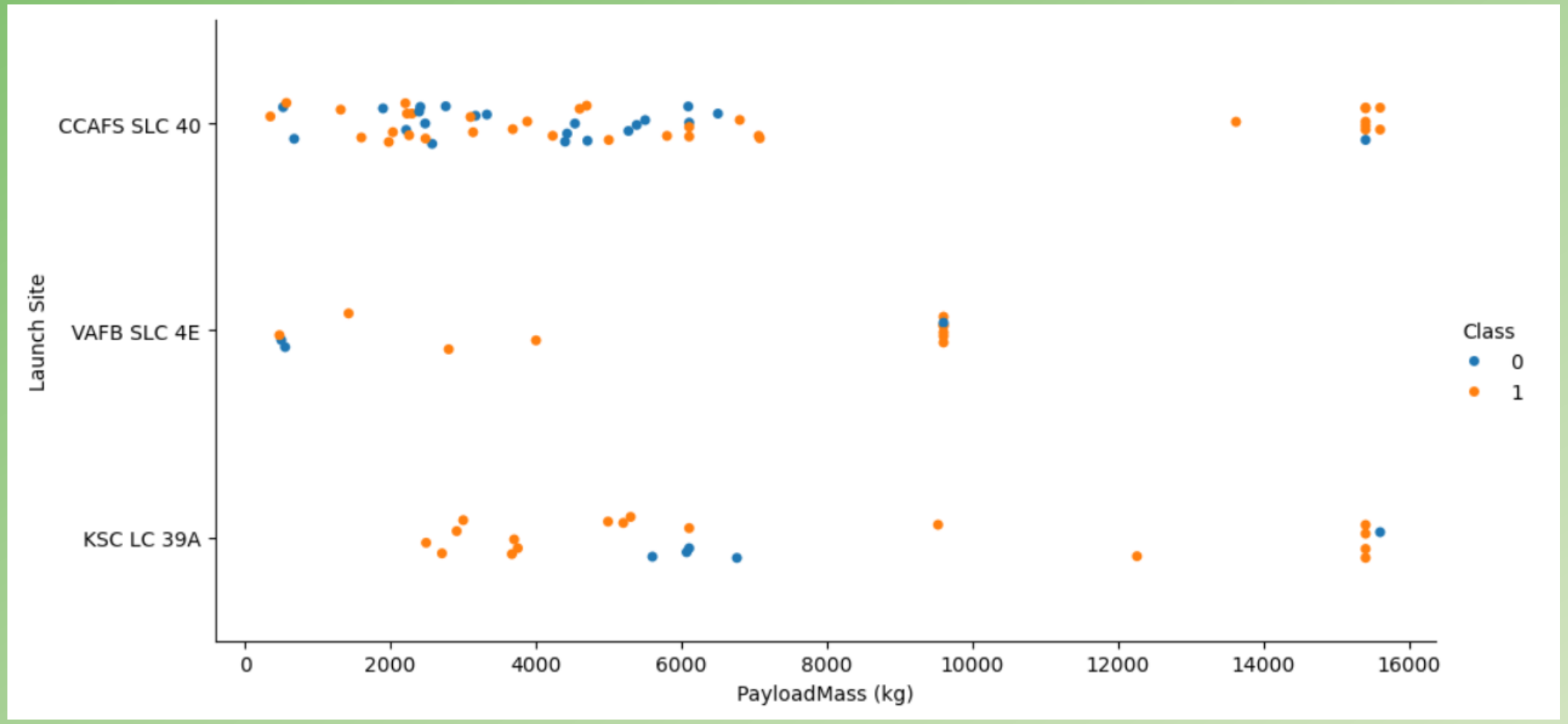


Flight Number vs. Launch Site



The success rate tends to increase with each new launch.

Payload Mass vs. Launch Site



Most launches have a payload under 7000 kg.
All payload launched from VAFB SLC-4E is under 10000 kg.

Success Rate vs. Orbit Type

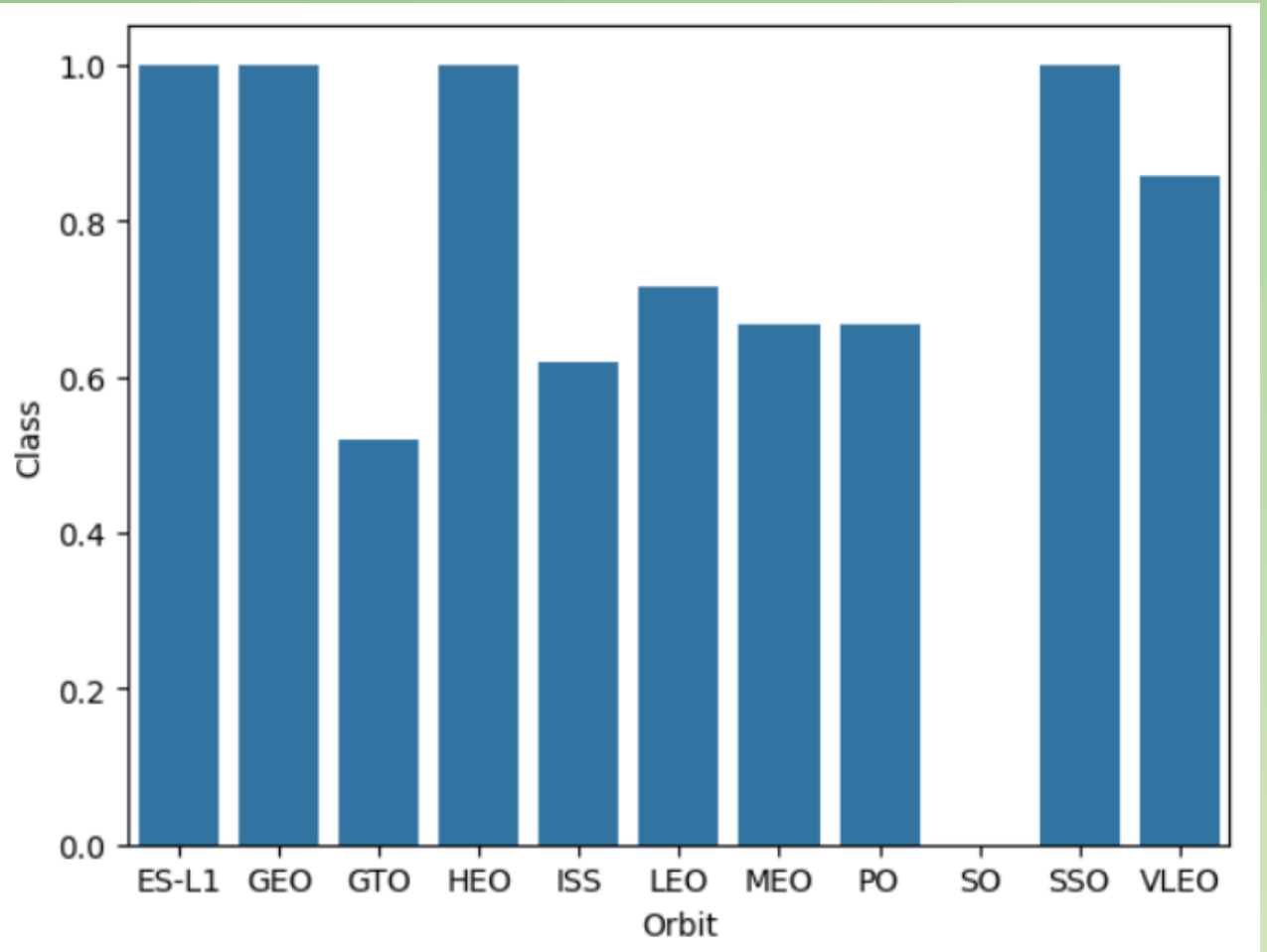
Four orbits have a 100% success rate:

ES-L1

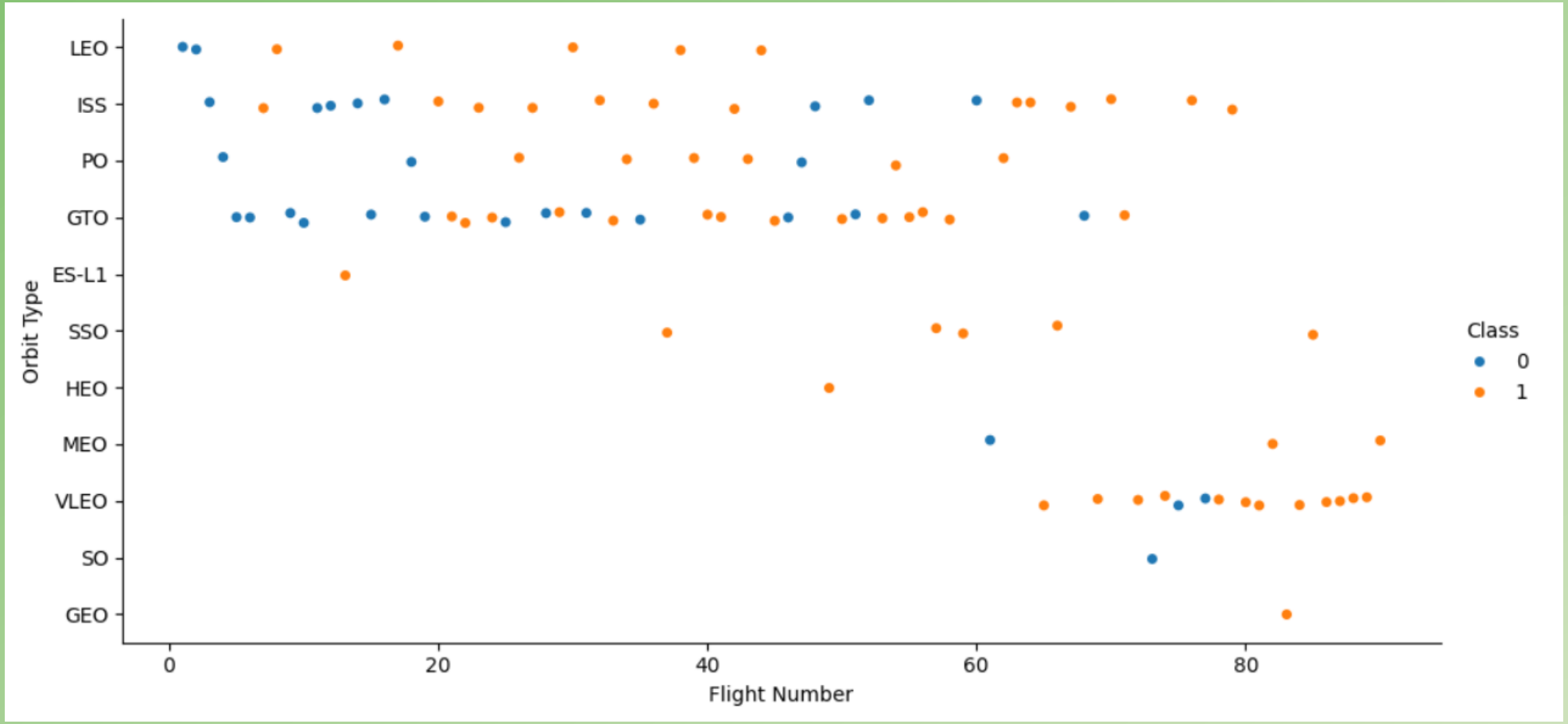
GEO

HEO

SSO

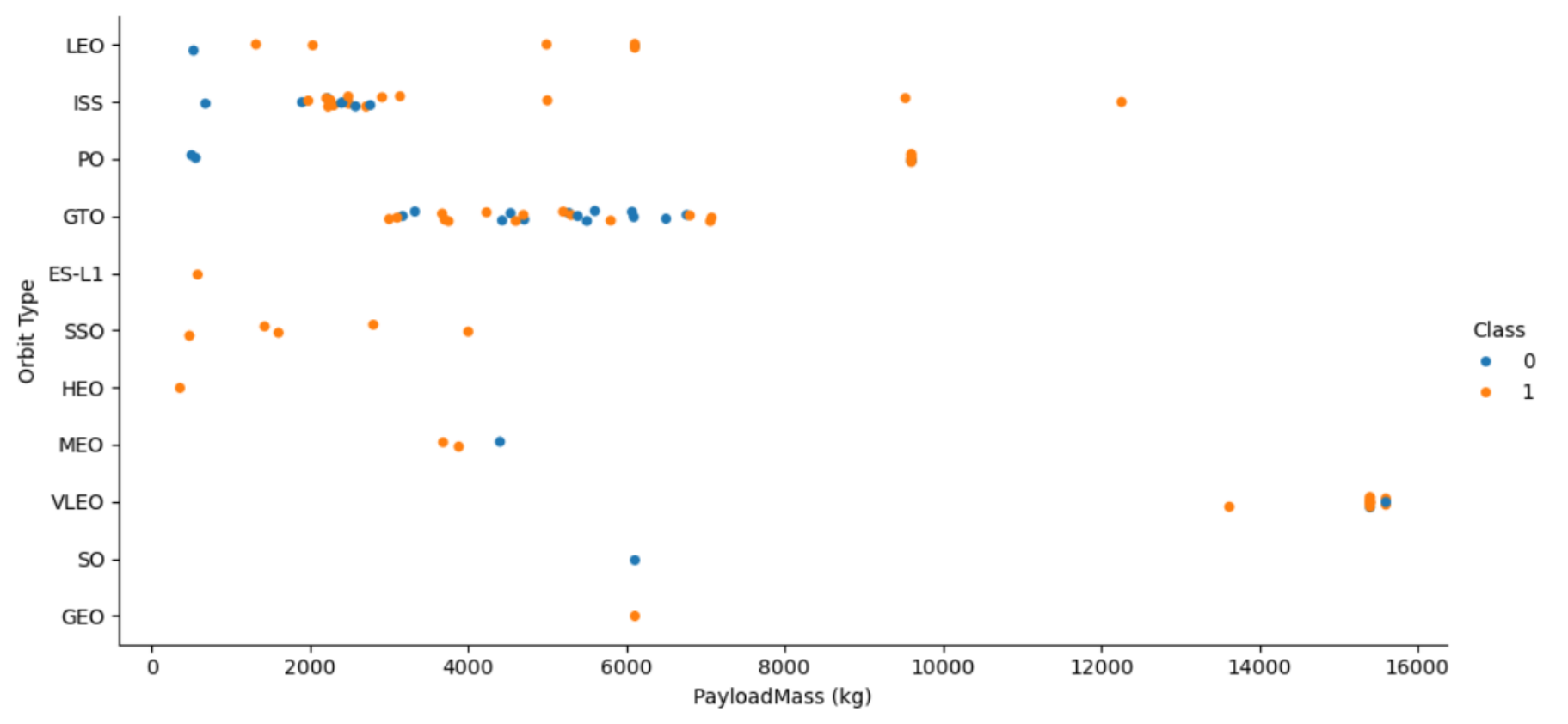


Flight Number vs. Orbit Type



A relatively high failure rate tends to be correlated with GTO and ISS orbits.

Payload Mass vs. Orbit Type

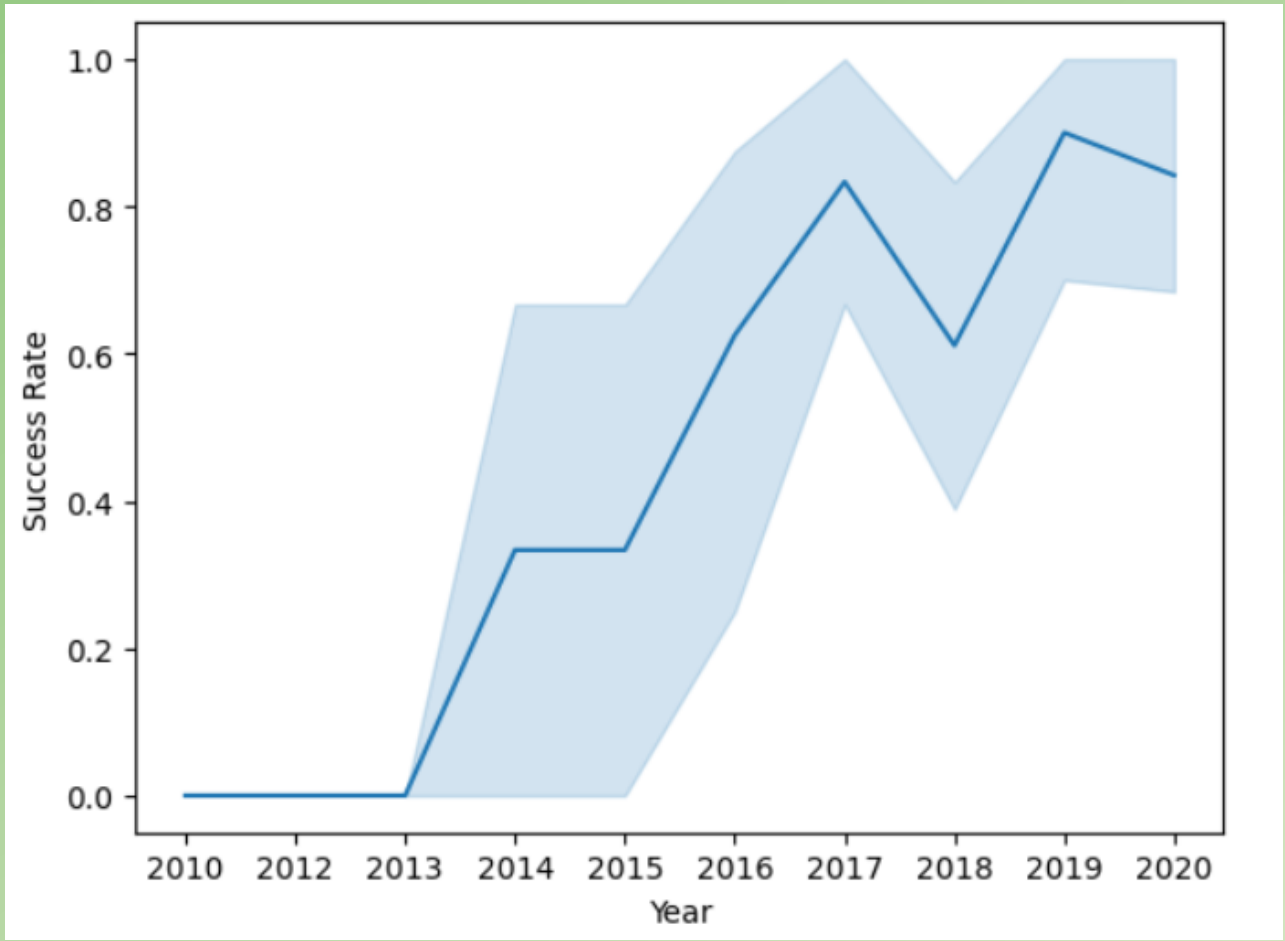


Very heavy payloads are often associated with successful missions to VLEO, ISS and PO orbits. ES-L1, HEO, LEO and SSO orbits tend to complete missions with relatively low-weight payloads.

Launch Success Yearly Trend

The number of successful launches has been increasing progressively since 2013, with a brief temporary decrease in 2018.

In recent years, the success rate surpassed 80%.



Finding All Launch Site Names

Query:

```
select distinct LAUNCH_SITE from SPACEXTBL;
```

Result:

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

The **distinct** keyword avoids returning repeated values.

There are four launch sites.

Launch Site Names Beginning with 'CCA'

Query:

```
select * from SPACEXTBL where LAUNCH_SITE like 'CCA%' limit 5;
```

Result:

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

With **like 'CCA%'**, the query returns records whose launch site name begin with 'CCA'.

Total Payload Mass for NASA

Query:

```
select SUM(PAYLOAD_MASS__KG_) from SPACEXTBL where CUSTOMER like 'NASA%CRS%';
```

Result:

SUM(PAYLOAD_MASS__KG_)
48213

The function **SUM** give us the total of the values in **PAYLOAD_MASS__KG_** column.
With **like 'NASA%CRS%'** we select only the records where the customer is NASA.

Average Payload Mass by F9 v1.1

Query:

```
select AVG(PAYLOAD_MASS__KG_) from SPACEXTBL where BOOSTER_VERSION is 'F9 v1.1';
```

Result:

AVG(PAYLOAD_MASS__KG_)
2928.4

The function **AVG** give us the average of the values in **PAYLOAD_MASS__KG_** column.
With **is 'F9 v1.1'** we select only the records where the booster model is 'F9 v1.1'.

First Successful Ground Landing Date

Query:

```
select MIN(DATE) from SPACEXTBL where LANDING_OUTCOME is 'Success (ground pad)';
```

Result:

MIN(DATE)
2015-12-22

MIN(DATE) give us only records whose date is the oldest and **where LANDING_OUTCOME is 'Success (ground pad)'** restrict the search to those where the landing outcome was successful.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

```
select BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME is 'Success (drone ship)' and PAYLOAD_MASS__KG_>4000 and PAYLOAD_MASS__KG_<6000;
```

Result:

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

This query finds the names of the boosters that succeeded in missions whose payload was bigger than 4000 kg but less than 6000 kg. It uses the **and** operator to assure several conditions are simultaneously true.

Total Number of Successful and Failure Mission Outcomes

Query 1:

```
select count(MISSION_OUTCOME) as Success from SPACEXTBL where MISSION_OUTCOME is 'Success';
```

Result 1:

Success
98

Result 2:

Failure
3

Query 2:

```
select count(MISSION_OUTCOME) as Failure from SPACEXTBL where MISSION_OUTCOME is not 'Success';
```

Two queries and a bit of Boolean logic where used to find the total number of successful and failed missions.

Which Boosters Carried The Maximum Payload?

Query:

```
select BOOSTER_VERSION from SPACEXTBL where PAYLOAD_MASS__KG_ is (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

Result:

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

With the help of a subquery that uses the keyword **MAX** to get the maximum payload, we find which versions of boosters carried the maximum payload on a mission.

It turns out that these are several versions of the **Falcon 9 Block 5** booster rocket.

2015 Launch Records

Query:

```
select substr(DATE,6,2) as Month, BOOSTER_VERSION, LAUNCH_SITE, LANDING_OUTCOME from SPACEXTBL where substr(DATE,0,5)='2015' and LANDING_OUTCOME = 'Failure (drone ship)';
```

Result:

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)

This query finds the failed landing outcomes of the year 2015, indicating the month, launching site and booster rocket version.

We need to use **substr(DATE, 6, 2)** to extract the month from the date and **substr(DATE, 0, 5)** to extract the year because SQLite does not support a better method.

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

Query:

```
select LANDING_OUTCOME, count(LANDING_OUTCOME) as Event_Count from SPACEXTBL
where (DATE between '2010-06-04' and '2017-03-20') group by LANDING_OUTCOME order by Event_Count desc;
```

Result:

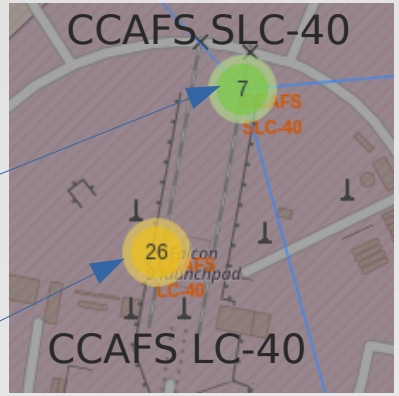
Landing_Outcome	Event_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

This query ranks the counts of each type of landing outcome from missions between June 4, 2010, and March 20, 2017, in descending order by quantity.

Interactive Visual Analytics

Launch Sites Map

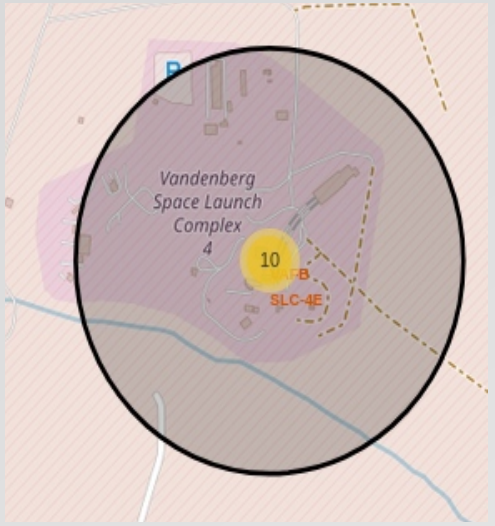
The four launch sites are located in the United States of America and are distributed across two main areas, one in California and one in Florida.



Launch Success Rate

A green mark indicates a successful launch while a red mark means a failure.

The 13 launches from KSC LC-39A site in California.



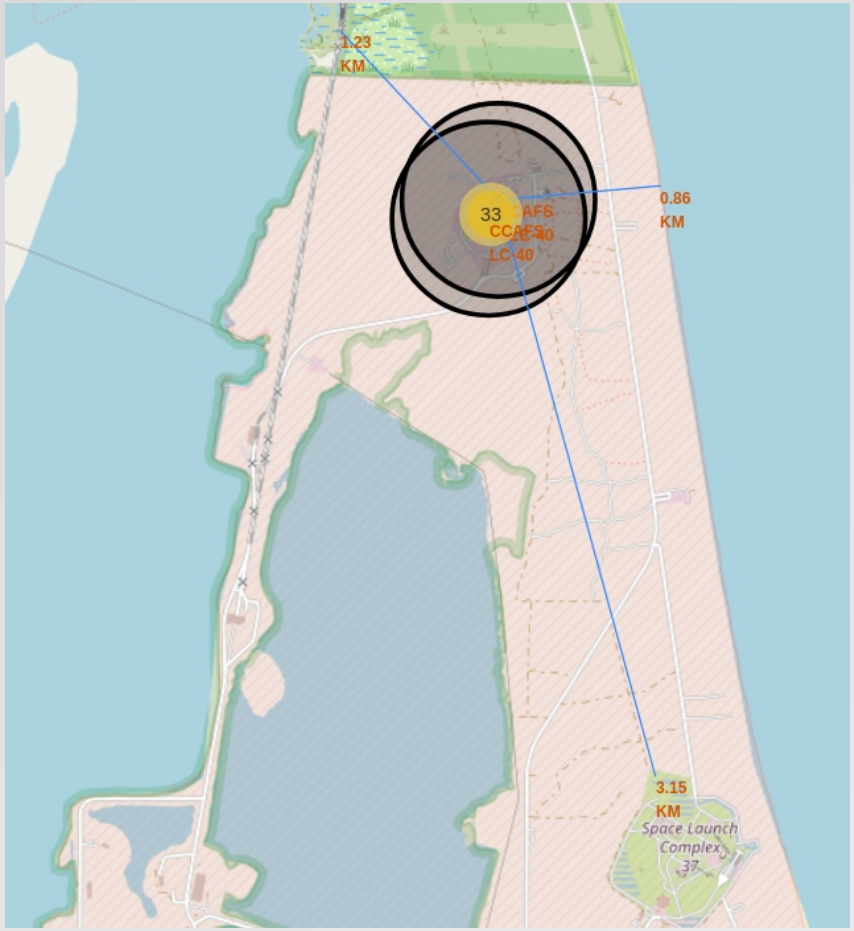
Outcome of each of the 10 launches at VAFB SLC-4E site in Florida.

Landmarks

All launch sites are located at a minimum distance of 15 kilometers from cities, which is obviously a decision that minimizes the risks of a crash near populated areas.

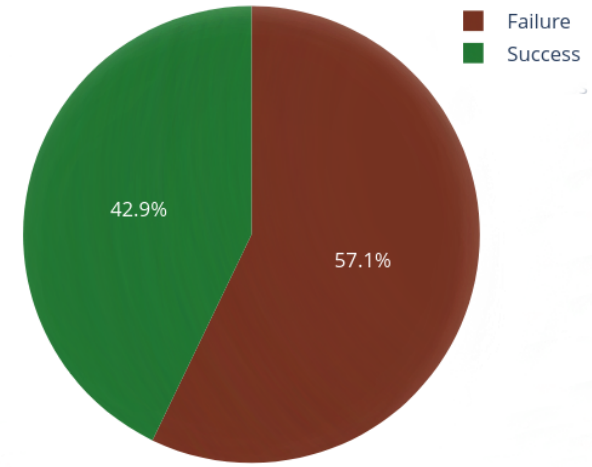
Since many rocket tests must land in the ocean, proximity to the coastline is very convenient.

There may be railways and roads nearby but these are always private and for transporting mission-related material to the launch site.

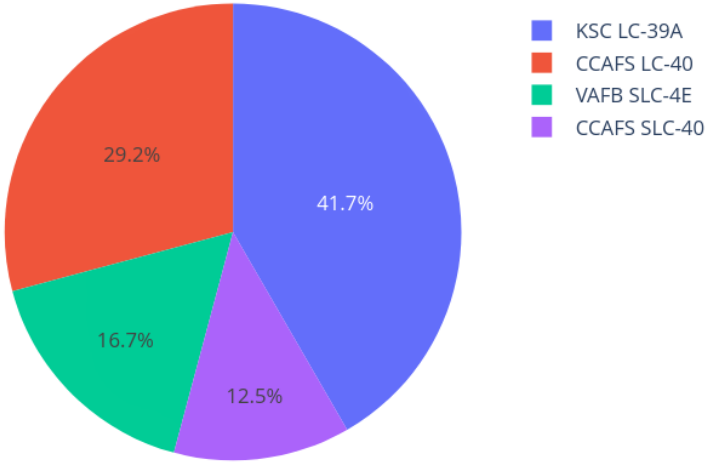


Surroundings of CCAFS SLC-40 and CCAFS LC-40.

Successful Launches by Site



All Successful and Failed Launches



Origin of successful launches

Only 42.9% of all the attempted launches were successful.

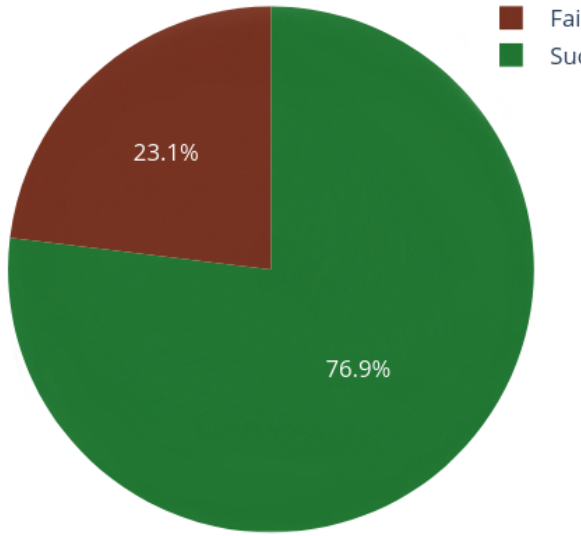
However, the rate of success vary by site.

A 41.7% of all successful launches were launched from KSC LC-39A.

Highest Success Outcome

Launch Site KSC LC-39A has the highest launch success rate (76.9%).

However, all launches from this site with a payload exceeding 5500 kg have failed.



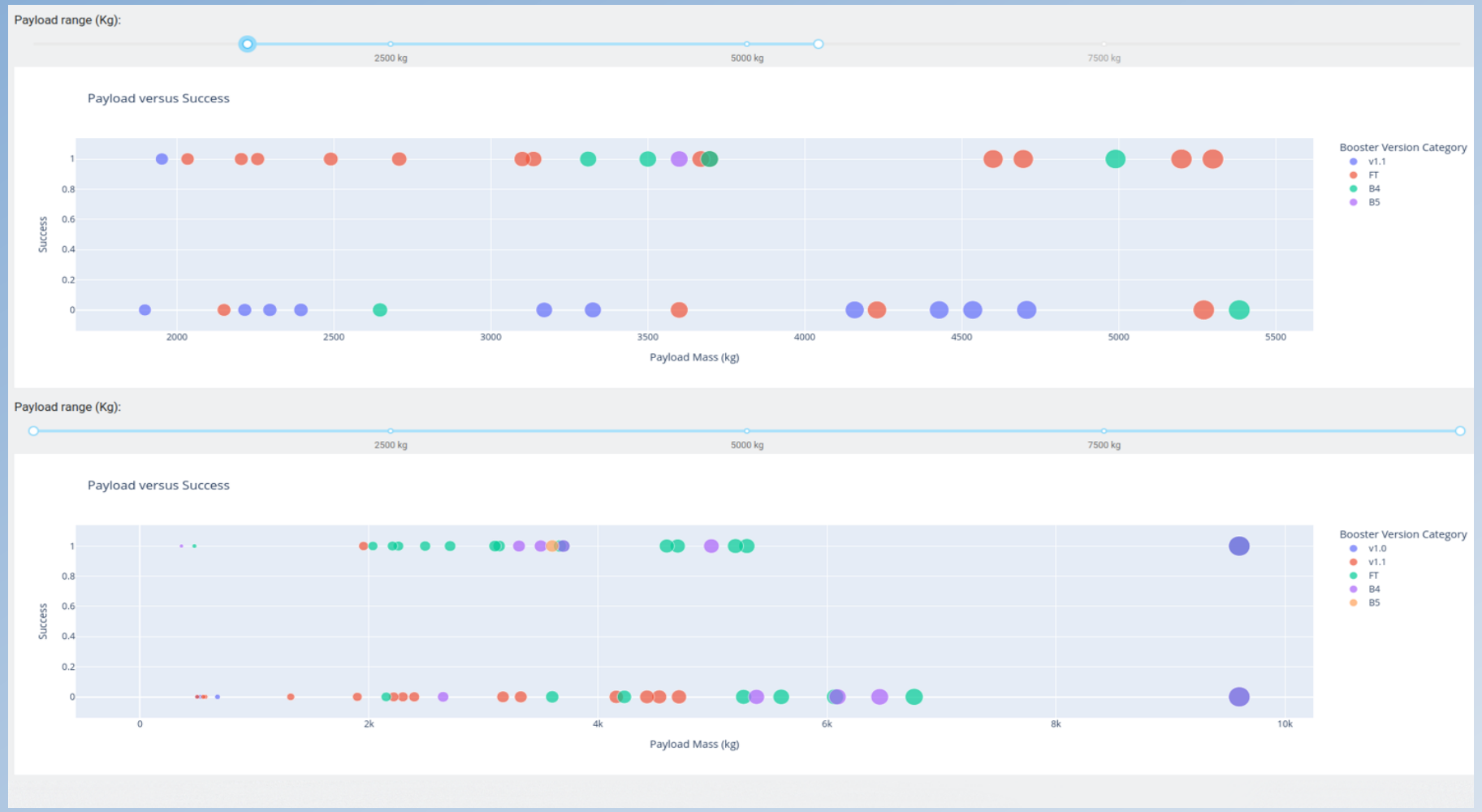
Launches from KSC LC-39A

Payload vs. Launch Outcome

Launches with a payload exceeding approximately 5500 kg are rarely successful.

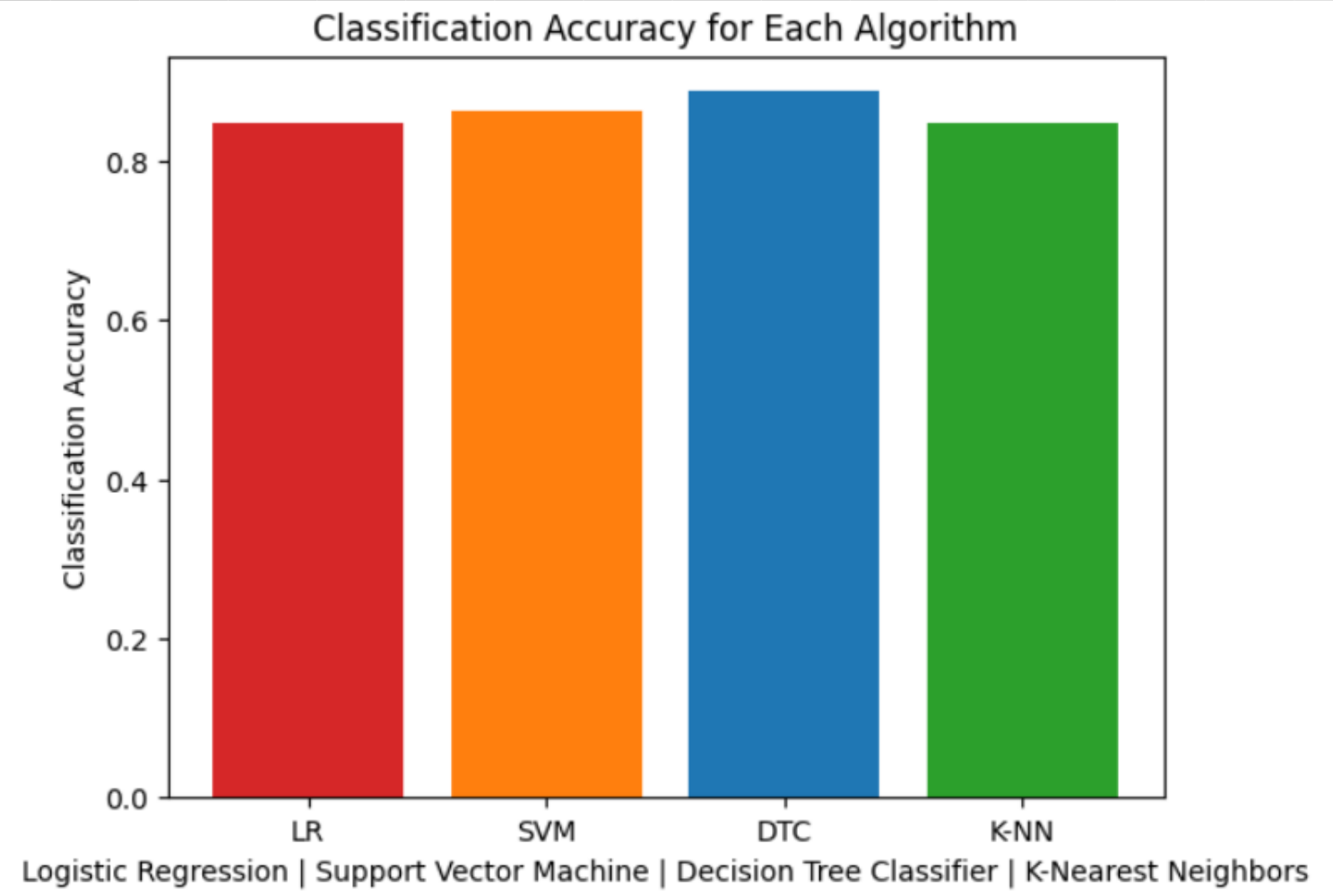
Optimal payload range appears to be between 1500 kg and 5500 kg.

The FT booster has the highest success rate and the B4 booster is the only one that has managed to complete a mission with the largest payload (9600 kg).



Predictive Analysis

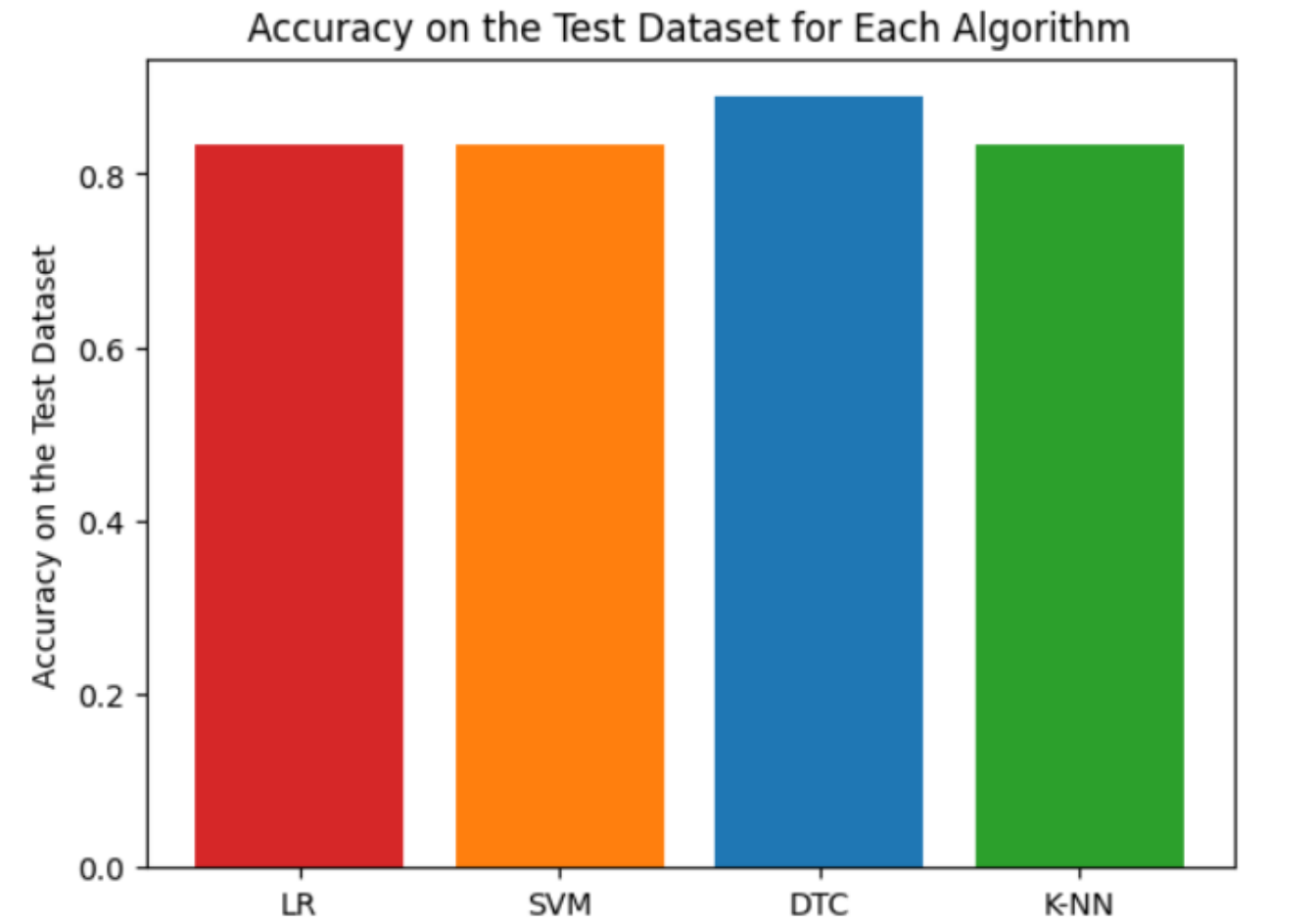
Classification Accuracy



Four ML algorithms were trained on the given data and the best one (by a small margin) was DTC (Decision Tree Classifier) with a slightly better classification score (**88.75%**) than SVM, which is a close second.

LR: 0.8482 | SVM: 0.8625 | DTC: 0.8875 | K-NN: 0.8482

Test Accuracy



The accuracy scores on the test data were the same for LR, SVM and K-NN models: **83.33%**

After iterating and tweaking the parameters a bit, DTC showed a better accuracy score on the test data: **88.88%**

A Jupyter notebook with all the python code, results, and tuned hyperparameters is linked in the appendix section of this report.

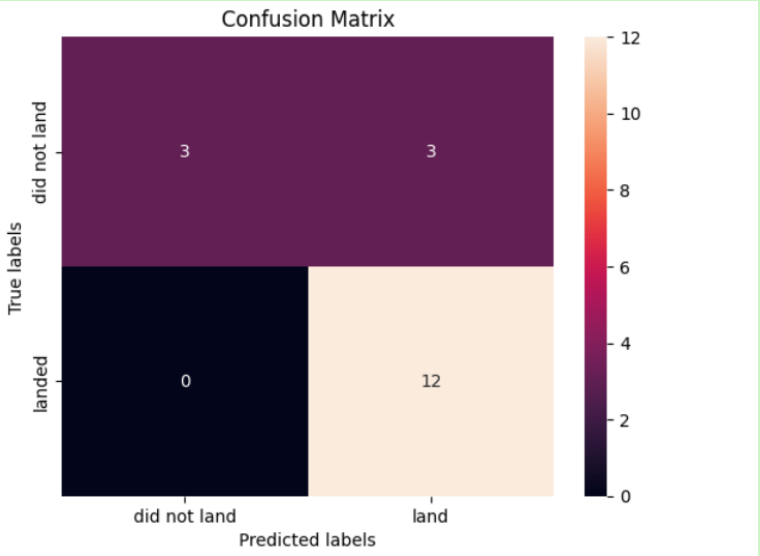
	Logistic Regression	Support Vector Machine	Decision Tree Classifier	K-Nearest Neighbors
Classification Accuracy	0.848214	0.862500	0.887500	0.848214
Accuracy on the Test Dataset	0.833333	0.833333	0.888889	0.833333

Confusion Matrix

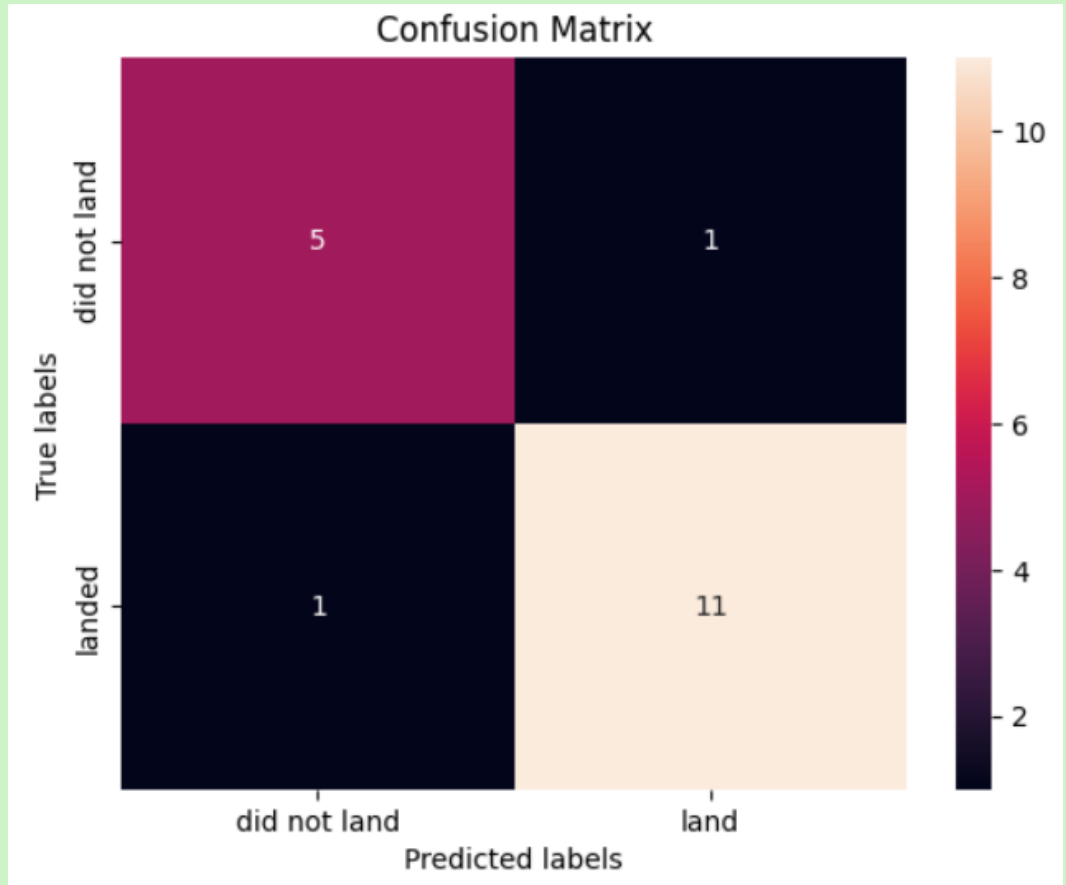
The confusion matrices of the LR, SVM and K-NN models are the same and show an accuracy of 83.3%.

However, the Decision Tree Classifier (DTC) model showed a better result and a confusion matrix with only a false positive and a false negative.

In the test data, DTC was correct **88.8%** of the time and is therefore the model of choice for future predictions.



Predictions made by LR, SVM and K-NN.



Decision Tree Classifier (DTC) predictions.

Conclusions

Conclusions

The success rate for SpaceX launches is increasing with time.

ES-L1, **GEO**, **HEO** and **SSO** orbits have the best rate of success.

Relatively small payloads tend to outperform heavy payloads.

A **41.7%** of all successful launches were launched from **KSC LC-39A**.

Several ML models were tested in order to find which one can predict the outcome of a launch more accurately.

The **DTC** model (Decision Tree Classifier) showed the best results on the test data, making 16 correct predictions of 18 (less than 12% of error).

However, as the model's classification accuracy is only 88.75% (only marginally better than the rest of the models), this result is likely due to a bit of stochastic luck and most of the time all models will perform equally (83.3% accuracy) given the small data sample used. Perhaps more data is needed to further improve these models.

Appendix

Python notebook of the ML predictive models and results:

https://github.com/MOOCsJunkie/SpaceX-Falcon9-First-Stage-Landing-Prediction/blob/main/SpaceX-Machine_Learning_Prediction.ipynb

Thank You!