IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Kostadin Slavov

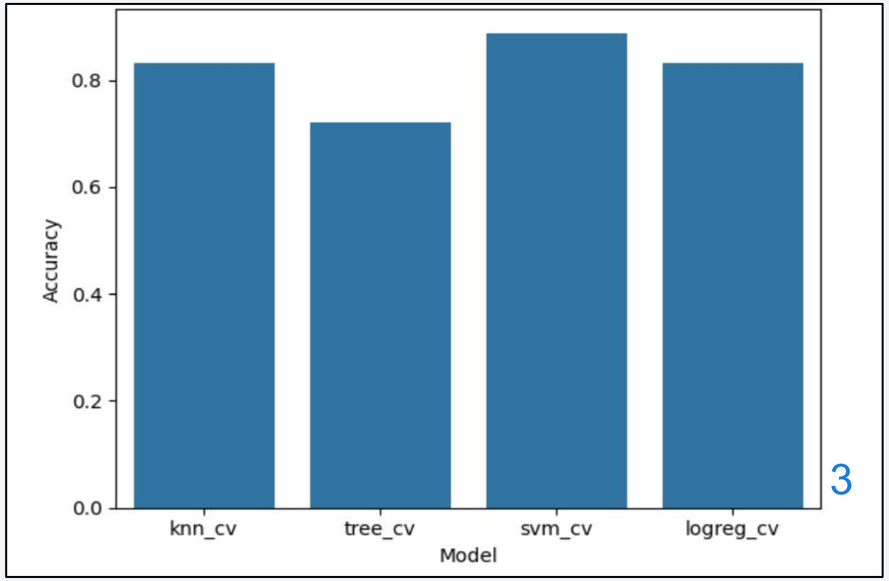
Jan 10, 2024

Outline

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

Executive Summary

- Summary of methodologies
  - We have used several techniques to address the problem and to predict whether or not a SpaceX rocket will land successfully or not.
- Summary of all results
  - Currently we can say with confidence of 83% that out support vector machine can predict whether the rocket will land successfully based on several parameters



Introduction

- The goal of our company is to find pattern in the mistakes of our competitor SpaceX, in order to prevent the failure of our rockets.
- The problems we want to find answers to are the following.
    - Which are the important parameters of the flight affecting the success rate.

Section 1

Methodology

Methodology

Executive Summary

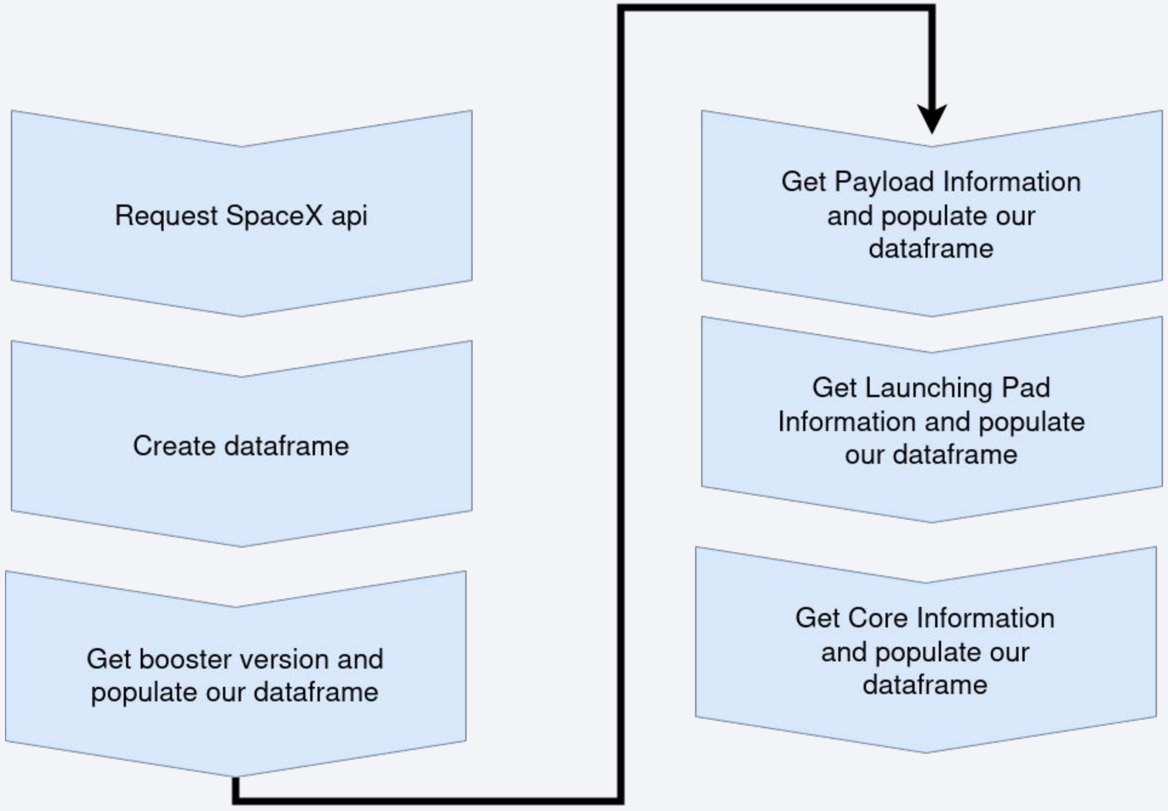
- Data collection methodology:
    - The data was collected from the [official SpaceX API]
- Perform data wrangling
    - Add a binary column class based on the Landing Outcome
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Standardize the data, split in training and test set, find the method that fits best for data predictions.

Data Collection

- We have used several sources of information
    - The spaceX api
    - Wikipedia

Data Collection – SpaceX API

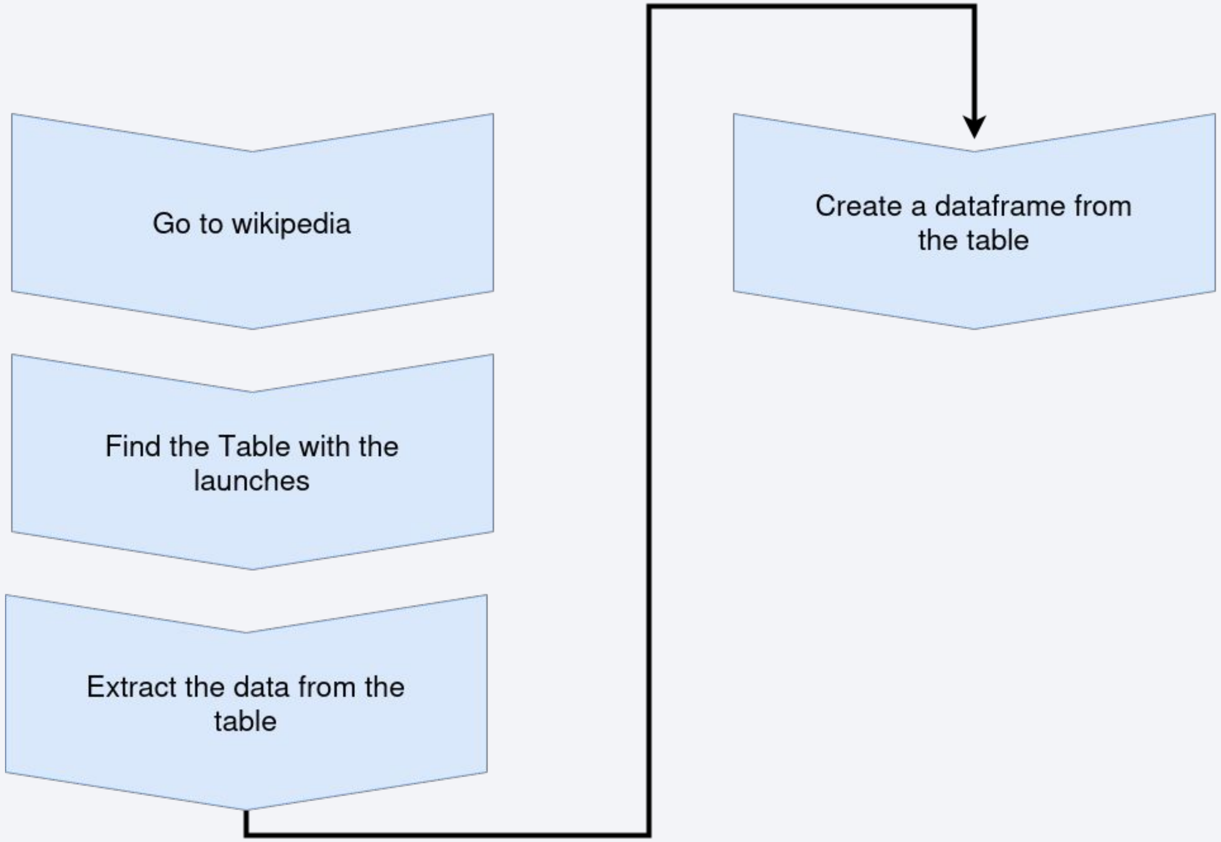
- Collect the data from official spaceX API.
- Get the Falcon 9 only data.
- Fill the missing values with mean in the PayloadMass column.
- Save the data.



- [Jupyter notebook for reference]

Data Collection - Scraping

- Find the data on Wikipedia
- Collect the data with web-scraping techniques
- Create a dataframe.



- [Jupyter notebook for reference]()

Data Wrangling

Analyze the data. Get to know and understand the data types, and calculate the following:

- Numbers of launches on each site
- Calculate the number and occurrences of each orbit
- Calculate the number and outcome of each orbit
- Create a binary class column for success and failure.

[Jypyter notebook for reference]

EDA with Data Visualization

Create different plots in order to observe if there is a pattern to be followed.

- PayloadMass vs FlightNumber
- LaunchSite vs FlightNumber
- PayloadMass vs LaunchSite
- Orbit vs Class
- Orbit vs FlightNumber
- PayloadMass vs Orbit
- Date vs Class

[Jupyter notebook for reference]

EDA with SQL

Performed sql queries:

- Get all of the LaunchSite
- Get launches from LaunchSite starting with “CCA”
- Get all of the launches where customer is NASA
- Compute the average Payload Mass for Falcon 9
- Get all of the Landing Outcomes
- Select the first Successful landing outcome
- Select All of the booster versions
- Count all of the Successful and Failed mission outcomes
- Get all of the boosters which carried the maximum payload
- Get all of the failures in 2015
- Get the landing outcomes for Failure (drone ship)/Success(ground pad) for a range of time

Jupyter Notebook for reference

Build an Interactive Map with Folium

The map includes several objects in order to visually represent the launching areas and their environments:

- Circle and marker for the NASA center
- Clusters of markers for each launching pad
- Distance between launching pad and a key point (Coast line, RailWay Highway)

Jupyter notebook for reference

Build a Dashboard with Plotly Dash

In the plotty dash we have added:
- Launch Site select
- Pie chart with success/failure ratio
- Payload slider

It is very interactive way to plot data based on launchpad and payload mass.

Project reference in GitHub

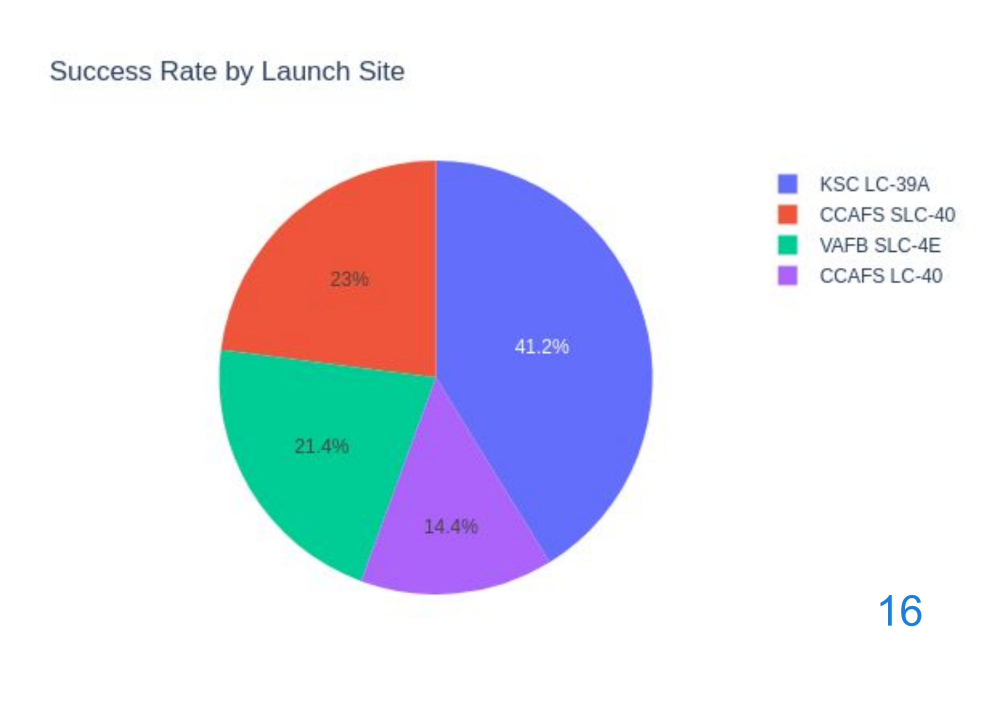
Predictive Analysis (Classification)

Performed actions during the Predictive Analysis

- Load the data and the transformed data
  - Transform the parameters for the prediction (X data)
  - Define the Y data and export it to numpy array
- Split the data in test and train with test size 0.2
- Perform the following classifications with features engineering in order to find the best model tuned with GridSearchCV
  - K-nearest neighbour
  - Decision Tree
  - Support Vector Machine
  - Logical Regression
- Create a confusion matrix as a result
- Compute the score of the models and their best parameters\
- Find the best fit model

Results

- We have come to conclusion that the different launching sites have different success rate.
- Some launching sites do not send massive payloads.
- Some of the orbits have great success rate than the others.

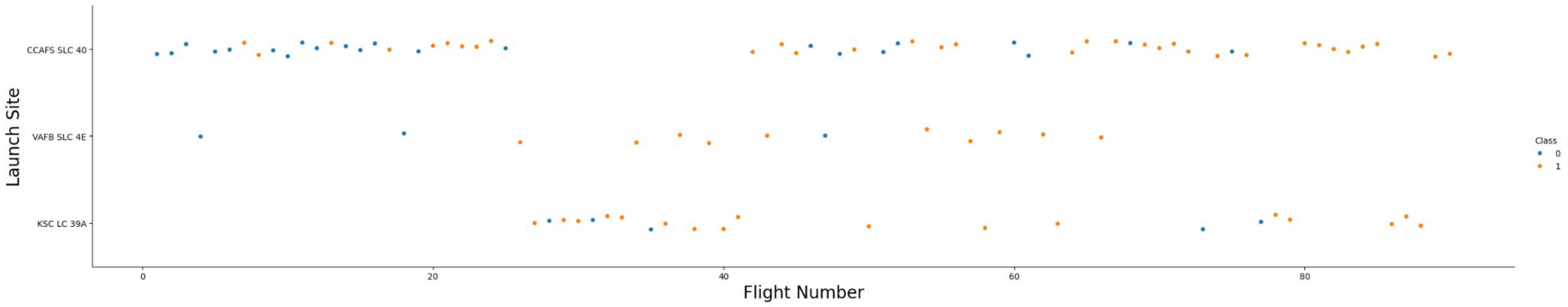


- Our most successful model was the SMV.

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

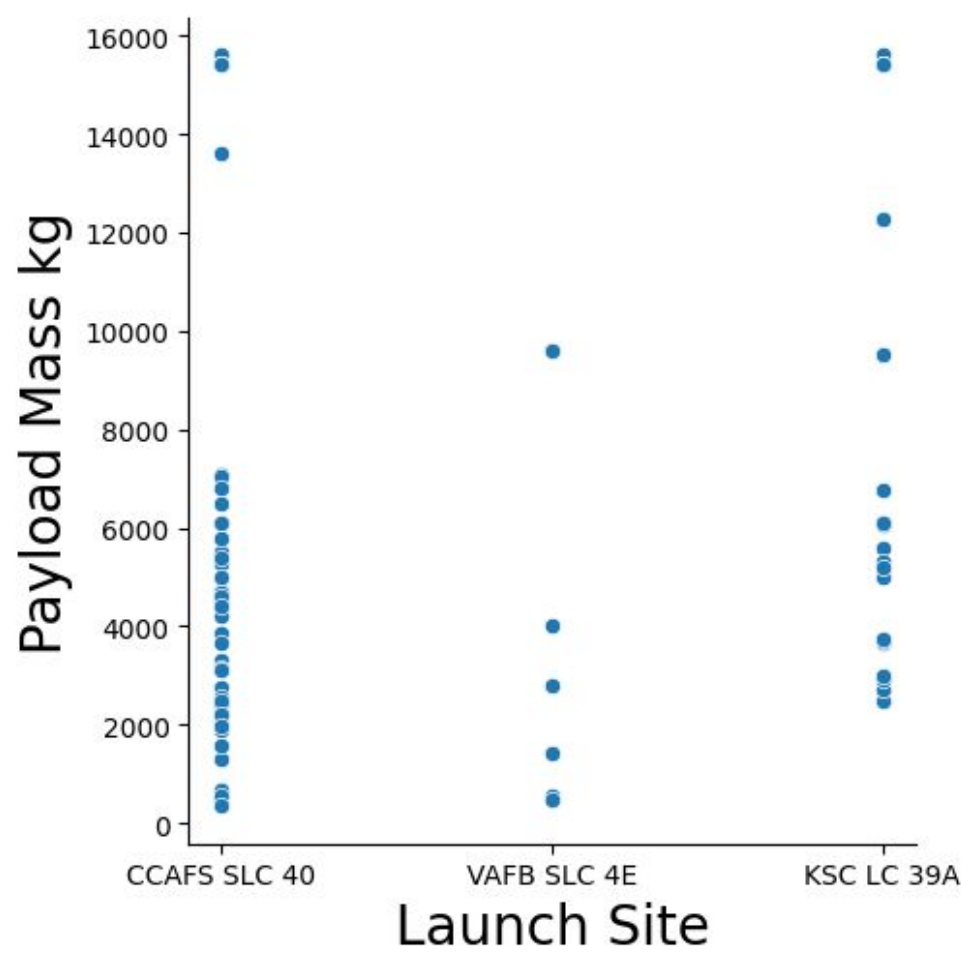


From this chart we understand that further in time the success rate improves which means that the company is taking care of the problems met during the failed ones.

Also VAFB SLC 4E has very low failure rate

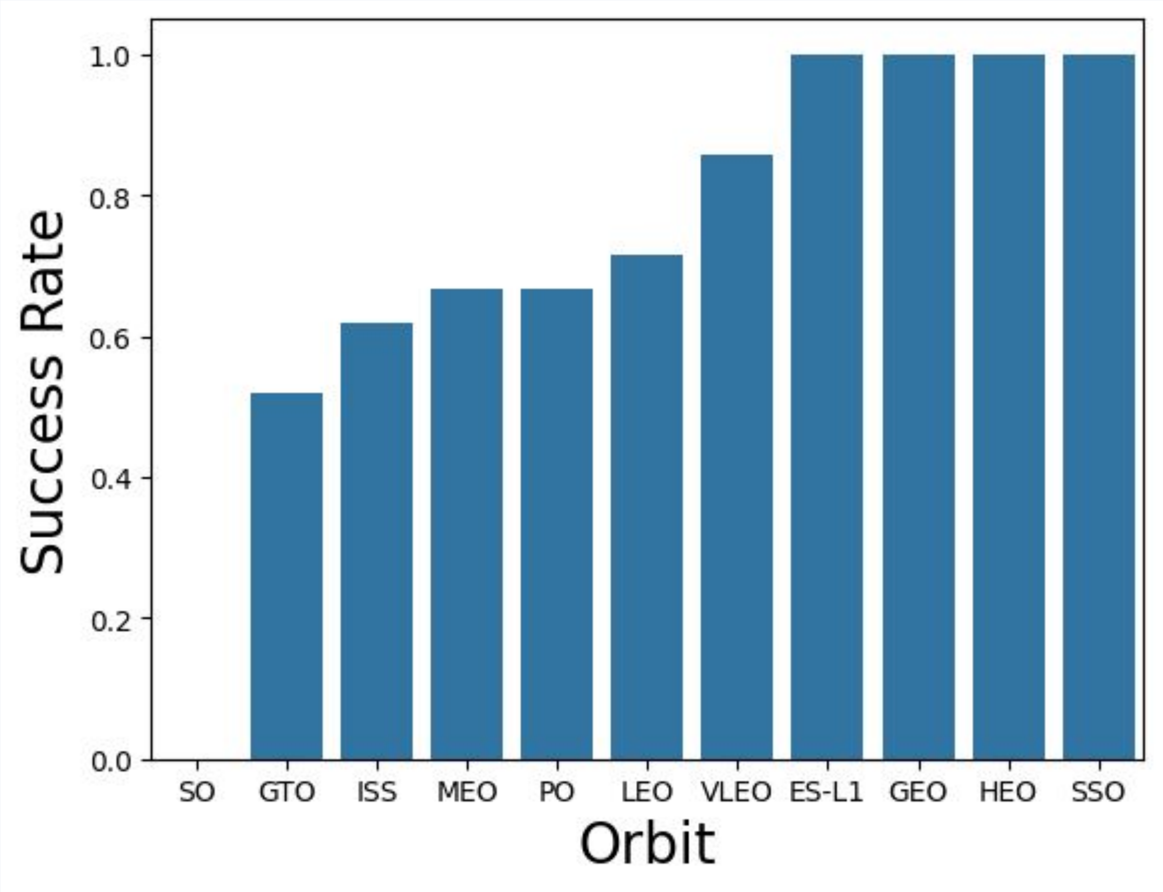
Payload vs. Launch Site

- VAFB SLC 4E never sent a payload larger than 10000kg. Which is why the success rate is so high.
- KSC LC 39A never sent a payload lighter than 3000kg.



Success Rate vs. Orbit Type

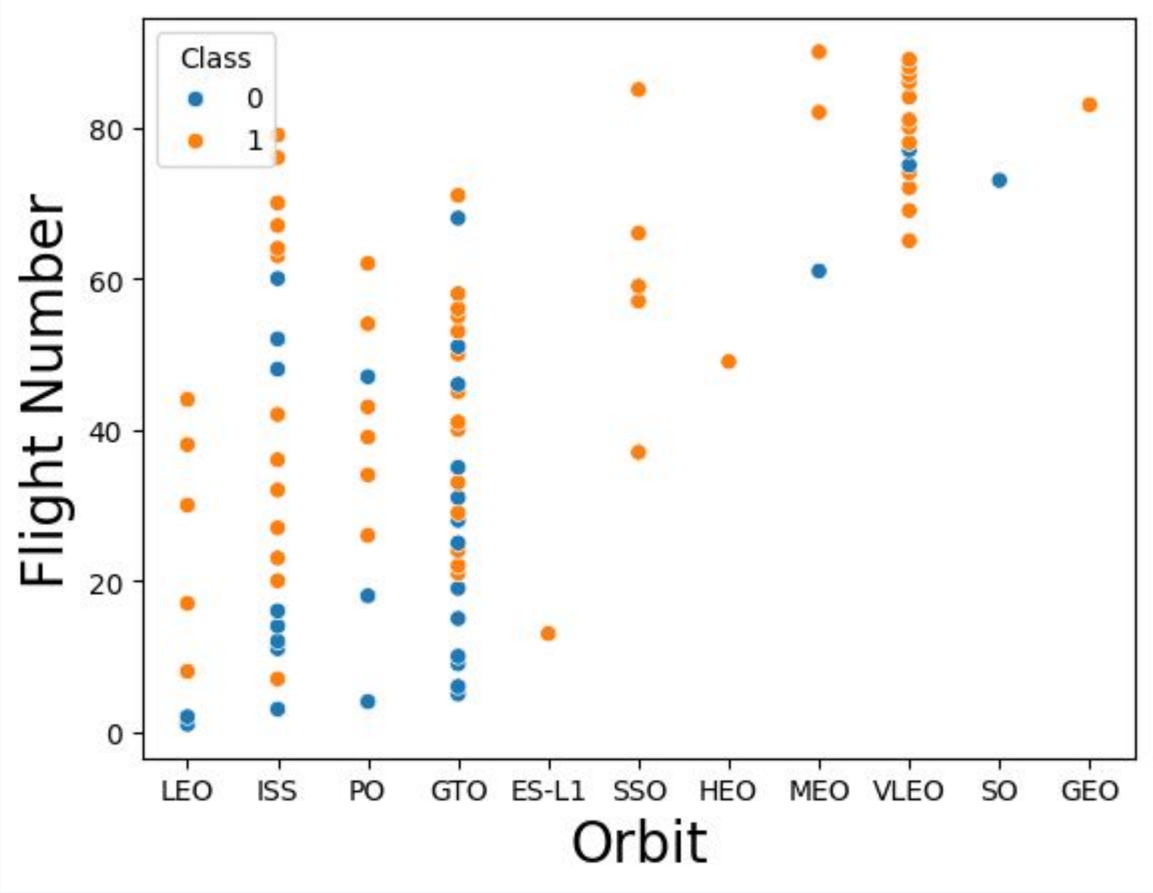
- We see that some of the orbits are dominating
- SO has success rate of 0



Flight Number vs. Orbit Type

With the flight number there is a positive trend for reaching higher orbits.

There is no weight of flight number to the higher obits because there is not enough data for the higher orbits yet.

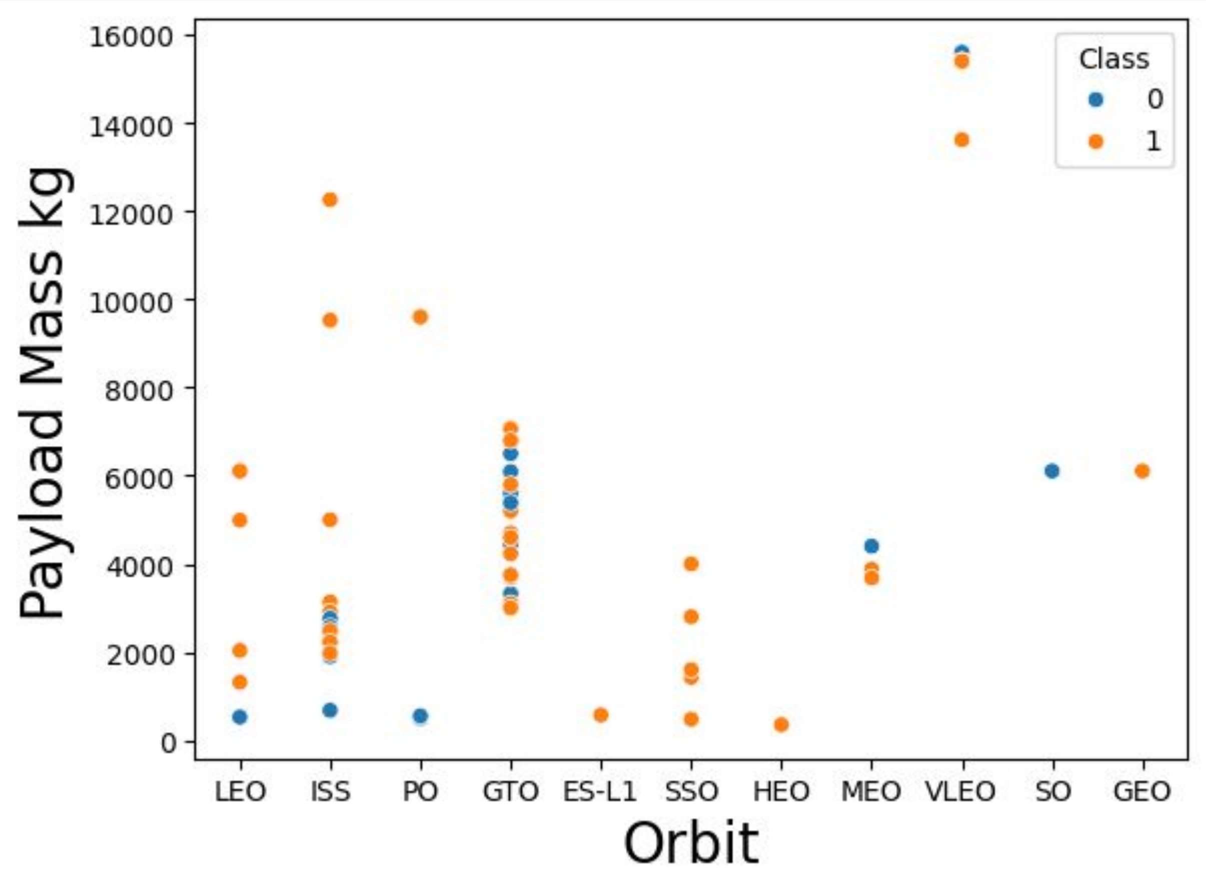


Payload vs. Orbit Type

There is only heavy payloads sent to the VLEO and 6000kg to SO and GEO.

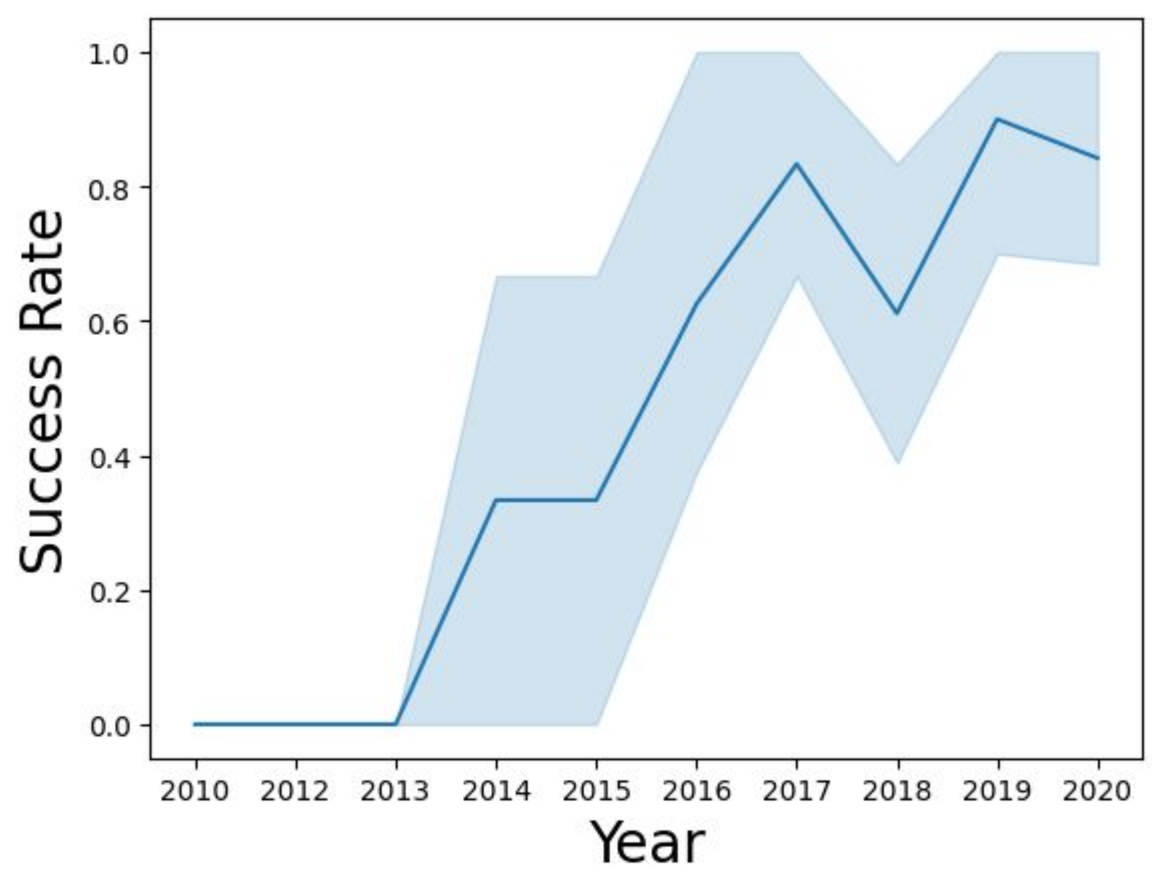
Most of the reached/attempted orbits are between 1000kg and 8000kg

SSO has 100% of success rate



Launch Success Yearly Trend

The success rate is increasing positively from 2013 until 2020 with a little low in 2018.



All Launch Site Names

We use the DISTINCT Keyword to get all of the values in Launch_Site column.

```
[11]: %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
 * sqlite:///my_data1.db
Done.
[11]:
```

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

Launch Site Names Begin with 'CCA'

We use the LIKE keyword to find them.

```
[14]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5;

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

[14]:

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

Total Payload Mass

We use the SUM keyword to sum the column PAYLOAD_MASS__KG_

```
[11]: %sql SELECT SUM(PAYLOAD_MASS__KG_) as payload_mass FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

[11]:

payload_mass
45596

Average Payload Mass by F9 v1.1

We use the AVG keyword and LIKE, because there are several subversions of the Falcon 9v.1.1%

```
[13]: %sql SELECT AVG(PAYLOAD_MASS__KG_) as AVG_PAYLOAD_MASS FROM SPACEXTABLE WHERE Booster_Version LIKE 'f9 v1.1%';

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

[13]:

AVG_PAYLOAD_MASS
2534.6666666666665

First Successful Ground Landing Date

Use the MIN keyword on the Date column to find the first successful ground landing date.

```
[14]: %sql SELECT MIN(Date)as first_successful_landing FROM SPACEXTABLE WHERE Landing_Outcome='Success (ground pad)';

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

[14]:

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

Use the range query and the distinct keyword to find all of the booster versions.

```
[44]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome="Success (drone ship)" AND 4000> PAYLOAD_MASS__KG_ < 6000 ;

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

[44]:

Booster_Version
F9 FT B1021.1
F9 FT B1022
F9 FT B1023.1
F9 FT B1026
F9 FT B1029.1
F9 FT B1021.2
F9 FT B1029.2
F9 FT B1036.1
F9 FT B1038.1
F9 B4 B1041.1
F9 FT B1031.2
F9 B4 B1042.1
F9 B4 B1045.1
F9 B5 B1046.1

Total Number of Successful and Failure Mission Outcomes

Use the COUNT keyword. We have observed that there is only one failed mission.

```
[59]: %sql SELECT COUNT(*) FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Suc%';

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

[59]:

COUNT(*)
100

```
[60]: %sql SELECT COUNT(*) FROM SPACEXTABLE WHERE Mission_Outcome LIKE 'Fai%';

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

[60]:

COUNT(*)
1

```
[62]: %sql SELECT COUNT(*) FROM SPACEXTABLE WHERE Mission_Outcome LIKE "Suc%" OR Mission_Outcome LIKE "Fail%";

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

[62]:

COUNT(*)
101

Boosters Carried Maximum Payload

Use subquery to find the max value of the payload mass, in order to perform the query.

```
[50]: %sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTABLE);

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

[50]:

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

Use substr as sqlite does not support filtering by Date.

```
[16]: %%sql SELECT substr(Date,6,2) as Month,
      Booster_Version,
      Landing_Outcome,
      Launch_Site
      FROM SPACEXTABLE WHERE substr(Date,0,5)="2015" AND Landing_Outcome LIKE "Fail%";
```

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

[16]:

Month	Booster_Version	Landing_Outcome	Launch_Site
01	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
04	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

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

We use again the substr for Date range and the BETWEEN keyword. Then we Order the results in descending order by Date.

[22]:
```
%%sql SELECT * FROM SPACEXTABLE
WHERE (Landing_Outcome="Failure (drone ship)" OR Landing_Outcome="Success (ground pad)")
AND substr(Date, 0, 11) BETWEEN '2010-06-04' AND '2017-03-20'
ORDER BY Date DESC;
```

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

[22]:

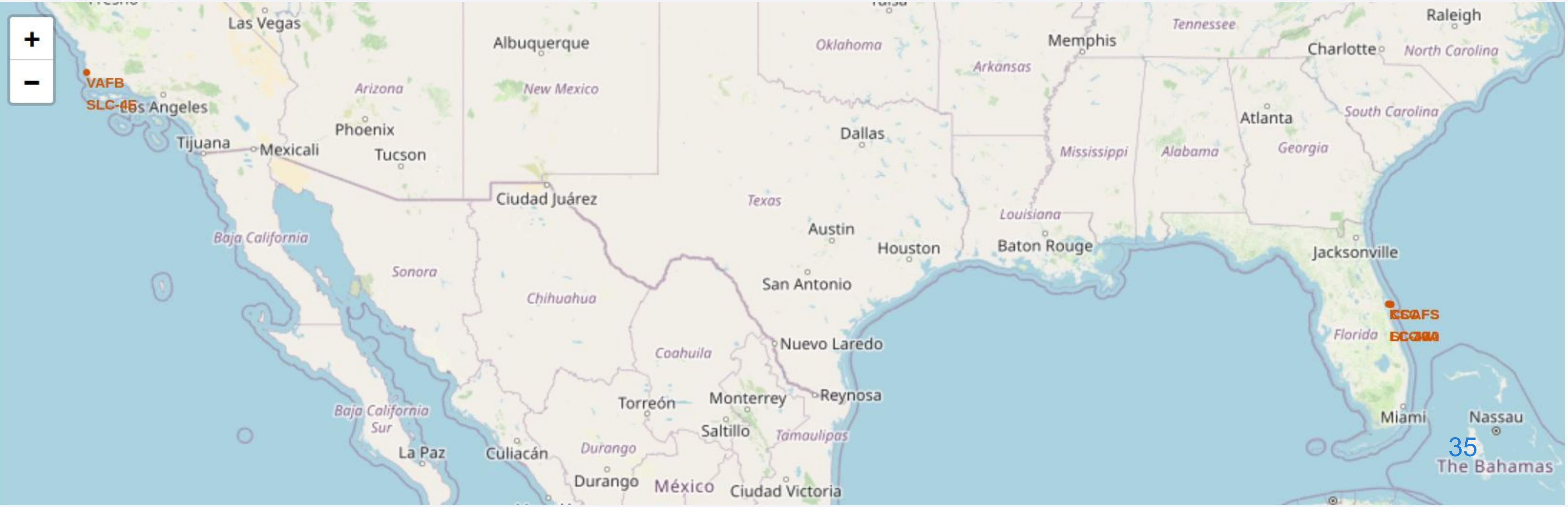
Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-07-18	4:45:00	F9 FT B1025.1	CCAFS LC-40	SpaceX CRS-9	2257	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
2016-06-15	14:29:00	F9 FT B1024	CCAFS LC-40	ABS-2A Eutelsat 117 West B	3600	GTO	ABS Eutelsat	Success	Failure (drone ship)
2016-03-04	23:35:00	F9 FT B1020	CCAFS LC-40	SES-9	5271	GTO	SES	Success	Failure (drone ship)
2016-01-17	18:42:00	F9 v1.1 B1017	VAFB SLC-4E	Jason-3	553	LEO	NASA (LSP) NOAA CNES	Success	Failure (drone ship)
2015-12-22	1:29:00	F9 FT B1019	CCAFS LC-40	OG2 Mission 2 11 Orbcomm-OG2 satellites	2034	LEO	Orbcomm	Success	Success (ground pad)
2015-04-14	20:10:00	F9 v1.1 B1015	CCAFS LC-40	SpaceX CRS-6	1898	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)
2015-01-10	9:47:00	F9 v1.1 B1012	CCAFS LC-40	SpaceX CRS-5	2395	LEO (ISS)	NASA (CRS)	Success	Failure (drone ship)

Section 3

Launch Sites Proximities Analysis

Launch Location sites

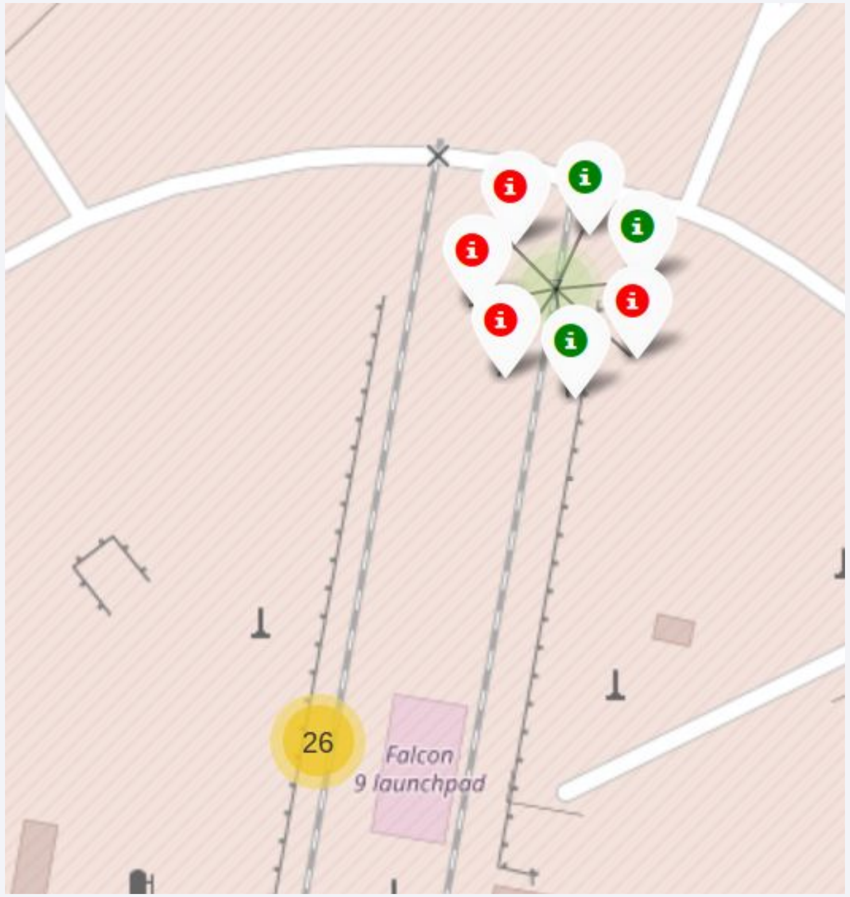
We can see that the launch location sites are very close to the East and West coast lines.



Clusters and success/failure markers

We can see that there is 46 launches from the East coast and 10 from the West coast.

Also we have added success/failure markers on the location.



Nearby Keypoints

In proximity of the launchpad there are:

- Coast line (1.3km)
- Railway (0.94km)
- Highway (7.65km)

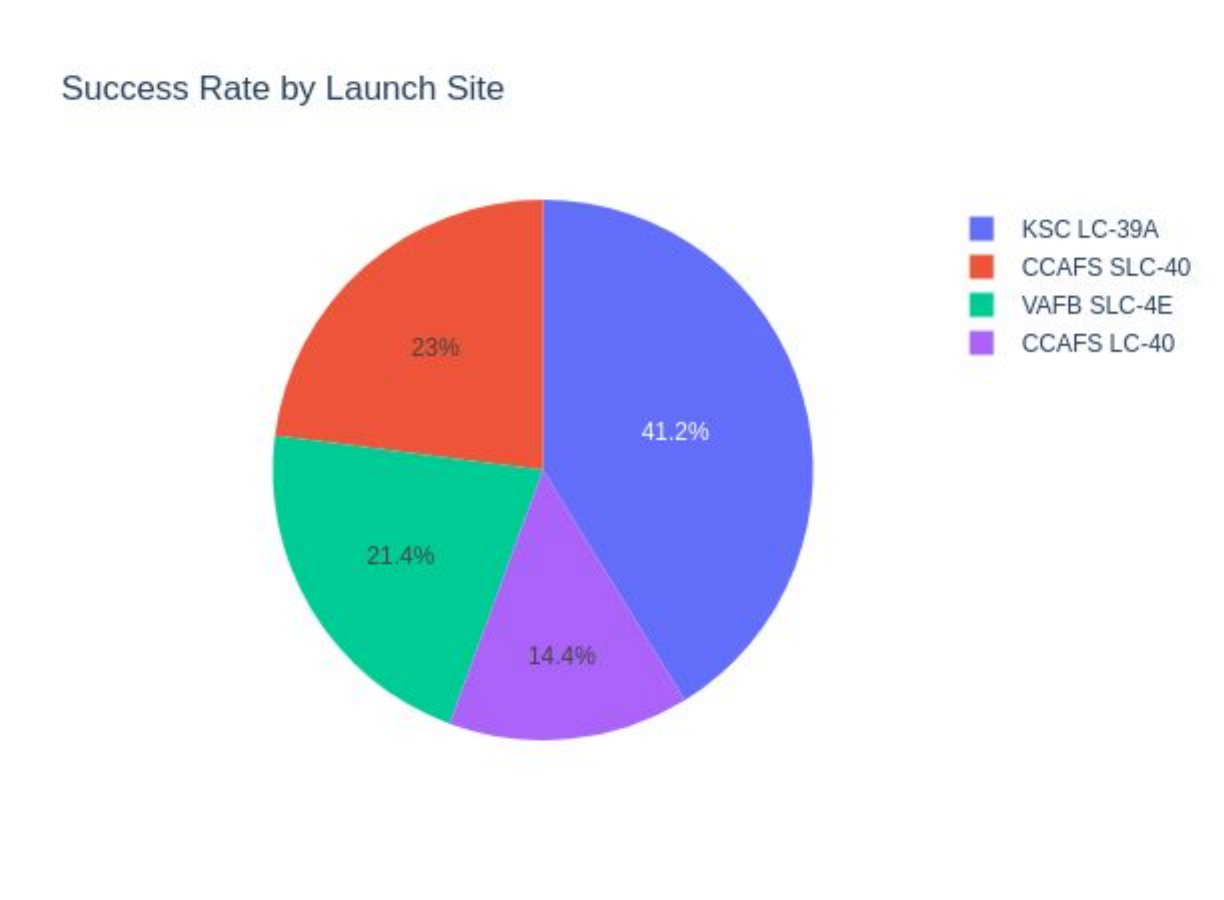
Section 4

Build a Dashboard with Plotly Dash

Interactive Dashboard Landing Success rate

KSC LC-39A is dominating with success rate of 41.2%.

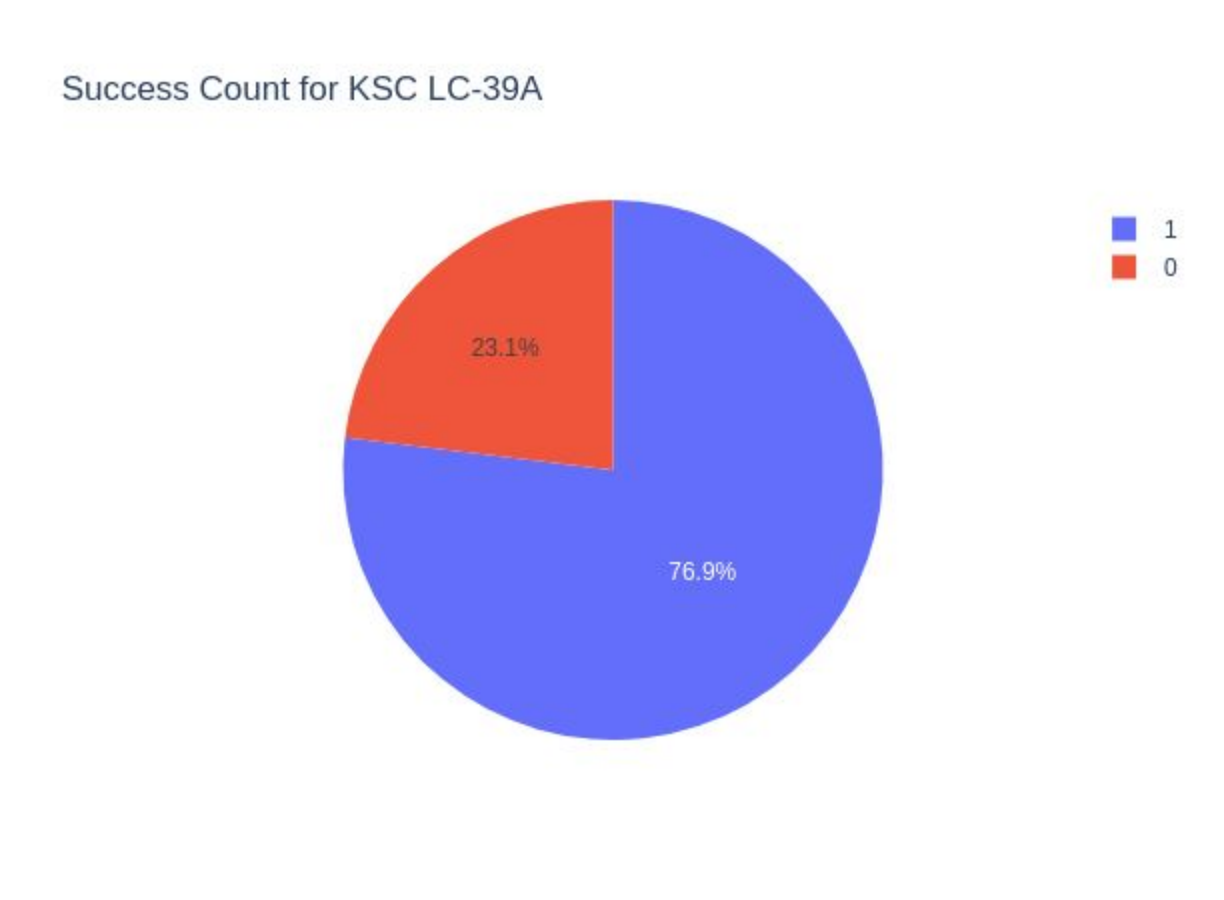
CCAFS LC-40 is doing very low percent wise.



Success Rate for KSC LC-39A

Only 23.1% of the missions have failed landing.

76.9% Success Rate



Light vs Heavy payloads

Up to 5000 kg

Most of the boosters carry under 5000 kg with a mixed percentage of success rate

Over 5000 kg

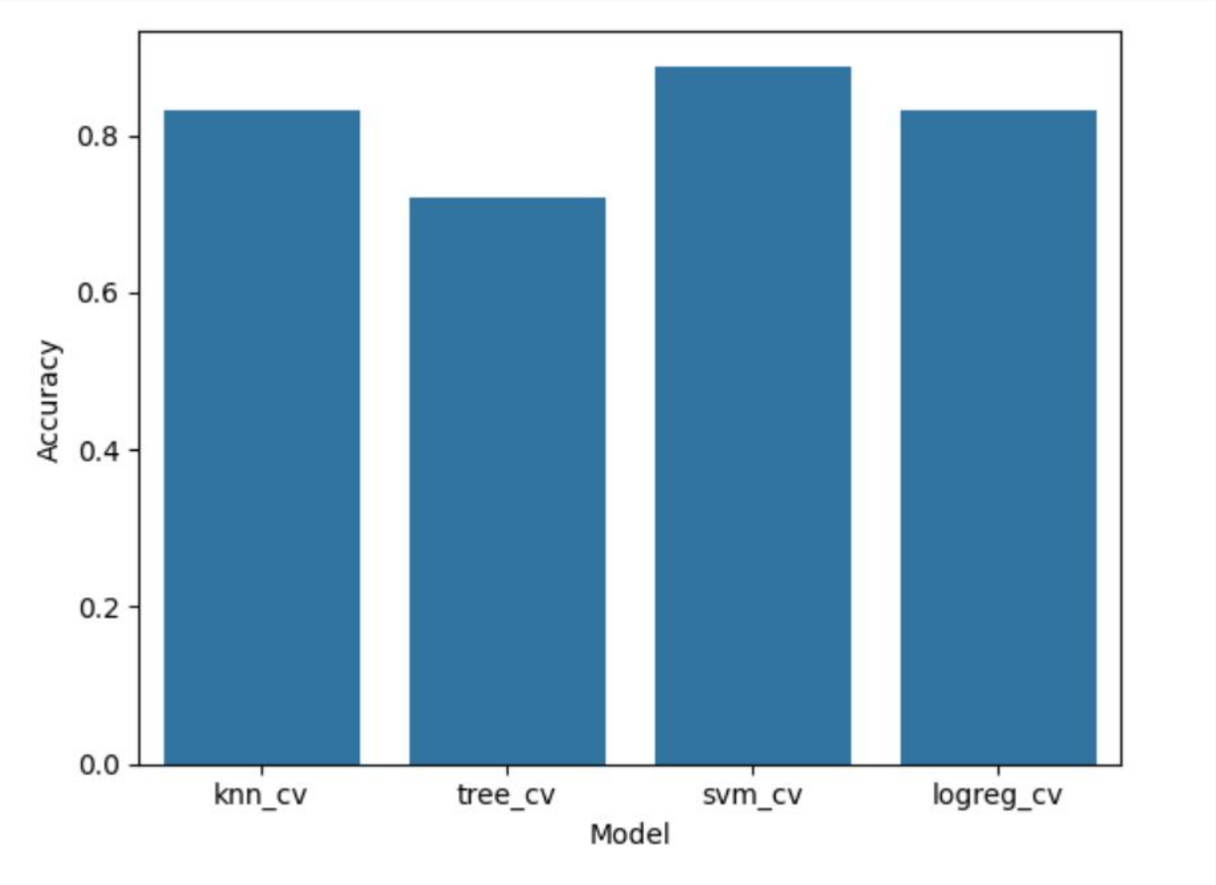
Fewer of the boosters carry over 5000 kg and the success rate drastically drops

Section 5

Predictive Analysis (Classification)

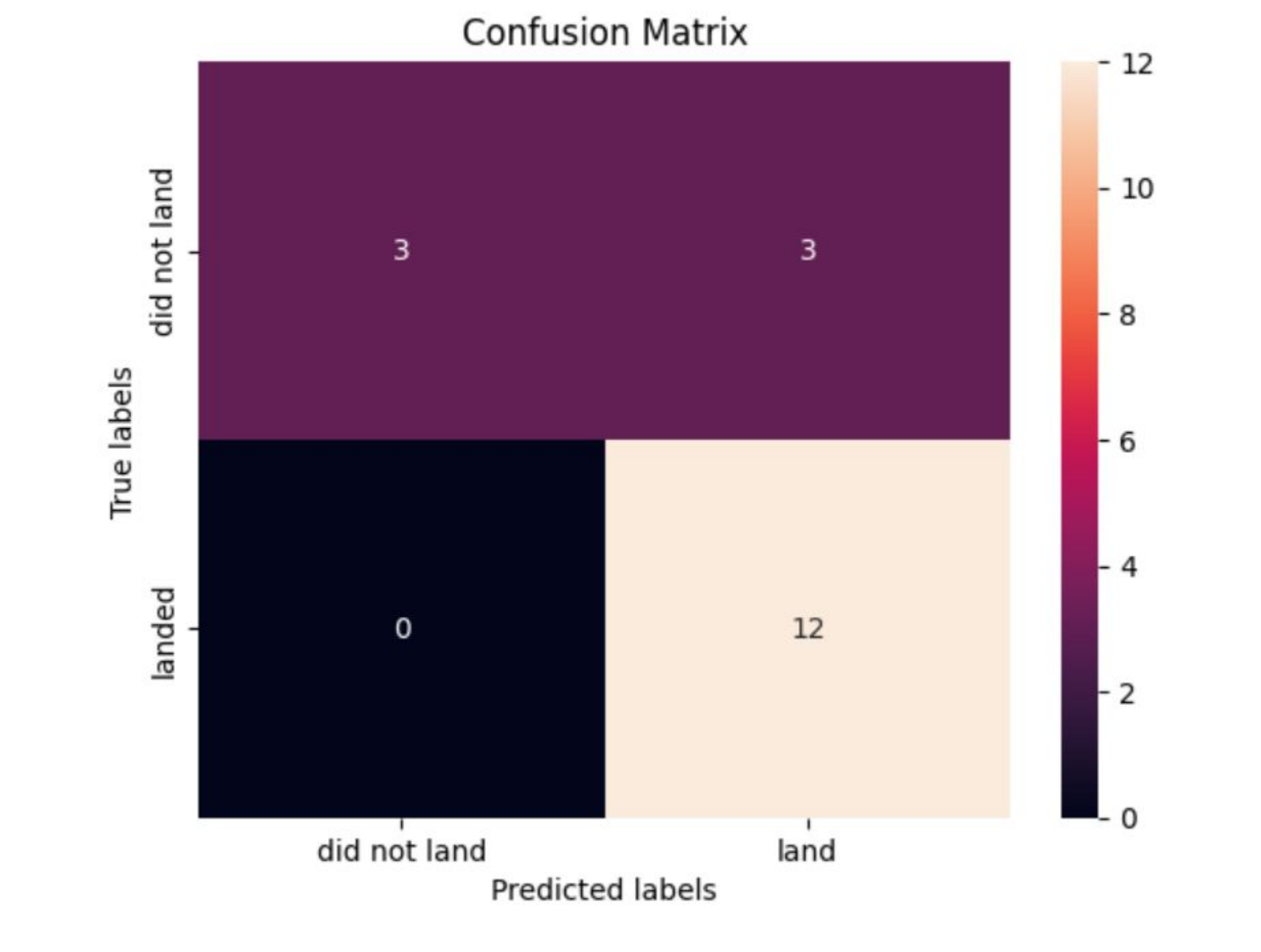
Classification Accuracy

- The model with highest accuracy is the Support Vector Machine with accuracy : 0.8482142857142856
- best parameters
  {
  'C': 1.0,
  'gamma': 0.03162277660168379,
  'kernel': 'sigmoid'
  }



Confusion Matrix

The model predicted correctly all of the landed cases, but it needs improvement with distinguishing certain negative cases.



Conclusions

- The success rate falls when the payload weight increases.
- ES-L1, GEO, HEO, SSO has the higher orbit success rate.
- Since 2015 the success rate is in positive trend.
- Each of the launch pads is close to the seashore, railway and highway.
- KCS-LC 39-A has the higher success.
- The average payload size is 2534kg.
- VAFB SLC 4E never sent a payload larger than 10000kg.
- KSC LC 39A never sent a payload lighter than 3000kg.
- Support Vector Machine is the best prediction model with confidence of 83%.

Appendix

All of the work can be found at [GitHub].

Thank you!