IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Colin Simpson
8 September 2023

Outline

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

Executive Summary

Data were collected on historical SpaceX Falcon 9 launches using the SpaceX API.

Exploratory data analysis was performed on the data to identify relationships between successfully landing the rocket first stage and the other mission features.

It was found that the payload mass, orbit type and flight date are related to the success rate. Machine learning models are able to accurately predict whether a mission will have a successful outcome in terms of landing the rocket first stage.

Introduction

SpaceX has a stated per launch cost of $62 million for their Falcon 9 rocket, significantly less than their competitors who typically cost upwards of $165 million.

SpaceX are able to offer such a low price due to their ability to recover the first stage of their rocket to be reused in subsequent missions. These recoveries are not always successful, so we are interested to know what features impact the likelihood of success.

We are also interested to know if we can predict the outcome based on the mission features, such as payload mass and orbit type.

Section 1

Methodology

Data Collection

Relevant data sets were collected from two separate sources:

- SpaceX Application Programming Interface (API).
- Web scraping an HTML table from Wikipedia.

Data Collection – SpaceX API

01	API call for Launch Data	• Flight number • Date • BoosterID, PayloadID, LaunchpadID, CoreID
02	API call based on BoosterID	• Booster name
03	API call based on PayloadID	• Payload mass • Orbit type
04	API call based on LaunchpadID	• Launchpad name • Launchpad latitude • Launchpad longitude
05	API call based on CoreID	• Landing success • Number of flights • Number of grid fins • Reused • Number of legs • LandingpadID

Data Collection - Web Scraping

01	**Obtain HTML page from URL**	• Obtain all webpage contents in HTML • Use BeautifulSoup object to contain this data
02	**Extract variable names from HTML table header**	• Extract all tables using <table> tag • Identify the desired table by inspection • Extract variable names using <th> tag
03	**Parse HTML table data to Pandas dataframe**	• Extract table elements using <td> tag and assign to variable using <th> tag • Convert extracted data to a Pandas dataframe

Data Wrangling

Collected data required some processing before it could be analysed:

- Filtered for Falcon 9 launches only.
- Five missing values for payload mass were replaced by mean of all other missions.
- There was no landing pad data for 26 of the 90 missions.
- A target variable was created to represent whether each launch was successful or not. This was determined based on the eight possible mission outcomes, such as “successful landing to ground pad” and “unsuccessful landing to drone ship”.

EDA with Data Visualization

Visualisations used as part of the exploratory data analysis (EDA):

- Categorical scatter plot of flight number and payload mass
- Categorical scatter plot of flight number and launch site
- Categorical scatter plot of payload mass and launch site
- Bar chart of success rate for each orbit type
- Categorical scatter plot of flight number and orbit type
- Categorical scatter plot of payload mass and orbit type
- Line plot of launch success by year

EDA with SQL - Part 1

SQL queries were used to answer some important questions, such as:

- Names of unique launch sites.
- Five records on launch sites that begin with “CCA”
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing on ground pad

EDA with SQL - Part 2

And additional questions like:

- Names of boosters with successful drone ship landings that had carried a payload between 4000 and 6000 kg.
- Total number of successful and failed mission outcomes.
- Names of booster versions that carried the maximum payload mass.
- Mission information for failed drone ship landings in 2015.
- Number of each landing outcome between 2010-06-04 and 2017-03-20.

Build an Interactive Map with Folium

An interactive map was built with Folium and included information such as:

- Markers and circles showing the name and location of the four SpaceX launch sites, to indicate their geographic setting.
- Coloured markers indicating the success/failure of each mission, collected into a cluster at the relevant launch site. This gives a visual indicator of the mission success rate at each launch site.
- A line and distance marker to a nearby coastline point from one launch site. This shows nothing of particular interest, but was a necessary component of the lab.

Build a Dashboard with Plotly Dash

We have created a dashboard that includes the following visualisations:

- Pie chart showing the proportion of successful missions across all launch sites.
- Pie chart showing the proportion of successful vs failed missions at each launch site.
- Scatter plot showing the payload mass vs outcome (successful or failed) at all sites or a specific site, filtered by the payload mass falling within a certain range.

These visualisations provides useful insights into the relative number of successful missions across the various launch sites used by SpaceX and how they relate to payload mass.

Predictive Analysis (Classification) - Part 1

The target variable of the predictive analysis is whether or not the first stage lands successfully. As this is a binary variable, the following machine learning algorithms, all of which are suitable for categorical prediction, were considered:

- Logistic regression
- Support vector machine
- Decision tree
- K nearest neighbours

Note that each column of data was normalised using the z-score, to ensure that no undue weighting was given to any particular feature.

Predictive Analysis (Classification) - Part 2

For each algorithm noted previously, the model was:

- Trained on a randomly allocated set of 80% of the available data, in which both the features and target variable are known.
- Tested on the remaining 20% of the available data, again with the target variable already known, to determine an accuracy score.

Where applicable, various different tuning parameters for a model were tested e.g. different kernel methods in the support vector machine algorithm. All four algorithms performed equally well, as will be shown later.

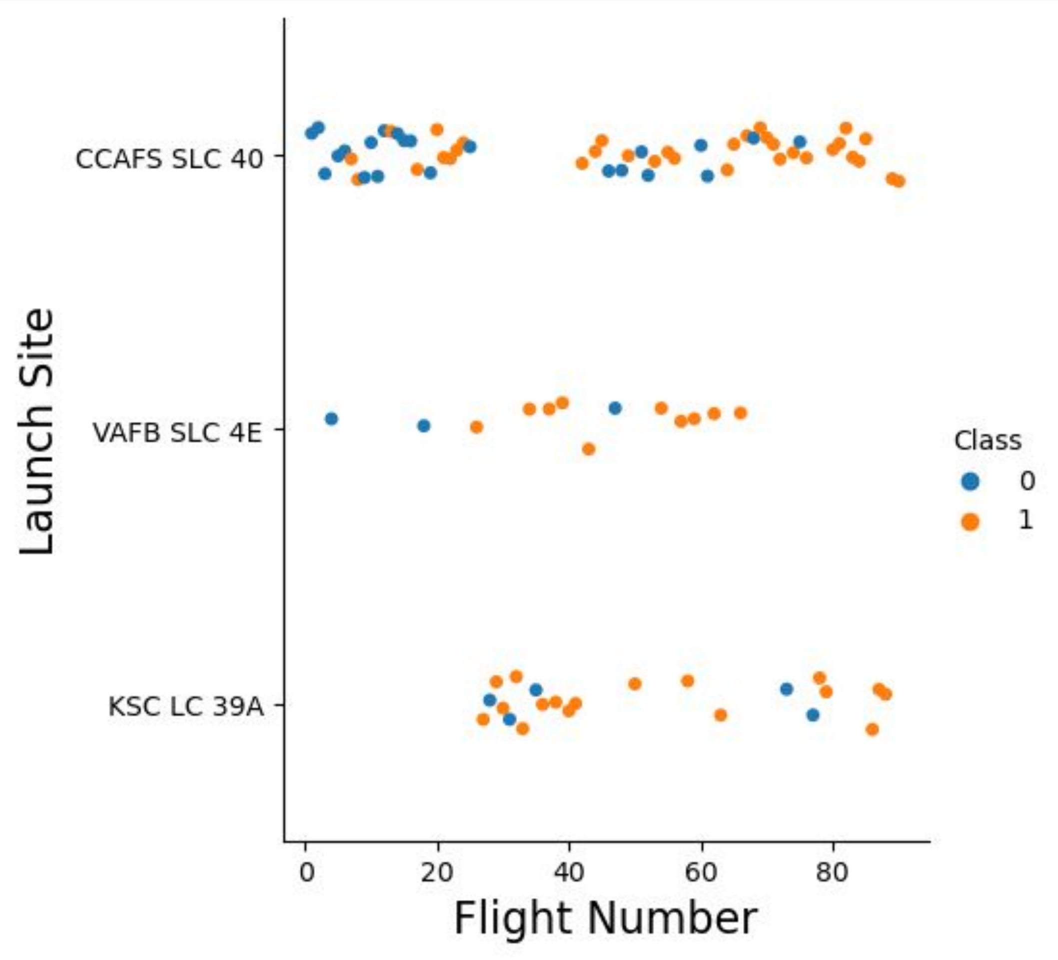
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

The scatter plot of Flight Number vs. Launch Site indicates that:

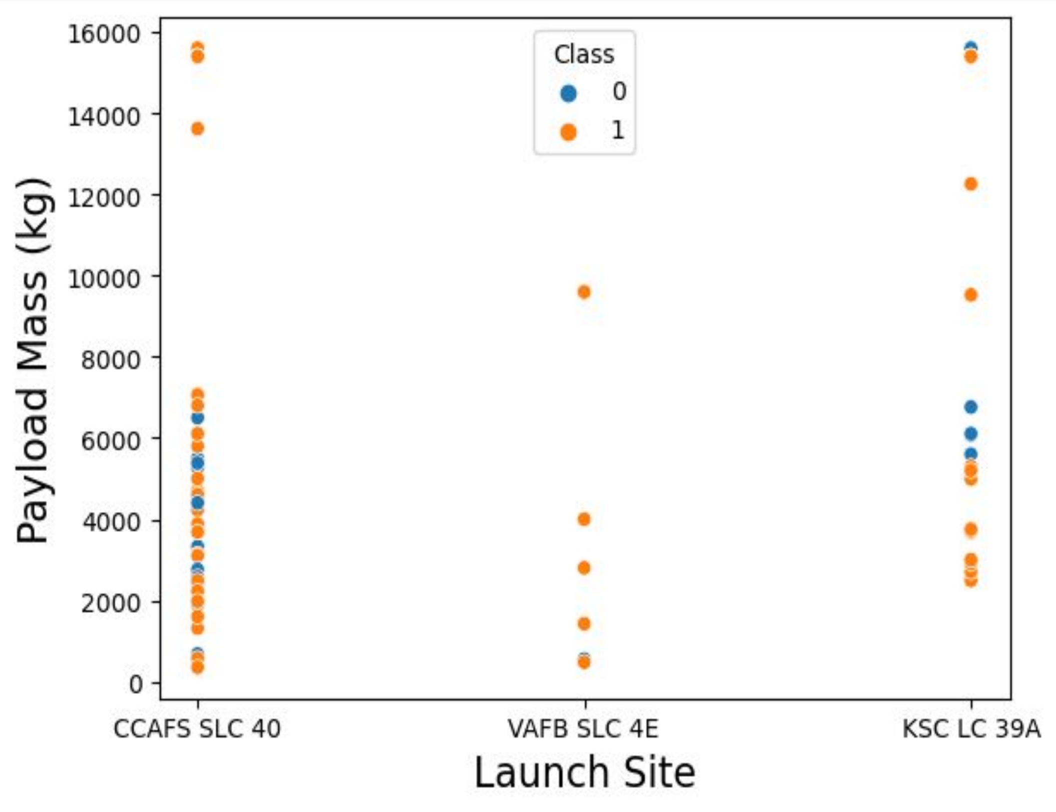
- Missions with higher flight numbers tend to be more successful (note: Class 1 means *successful*).
- Launch site CCAFS SLC 40 has the lowest success rate, particularly for lower flight numbers i.e. earlier dates.



Payload vs. Launch Site

The scatter plot of Payload Mass vs. Launch Site indicates that:

- VAFB SLC 4E does not have any launches with a payload mass over 10,000 kg.
- CCAFS SLC 40 has a wide range of payload mass launches.

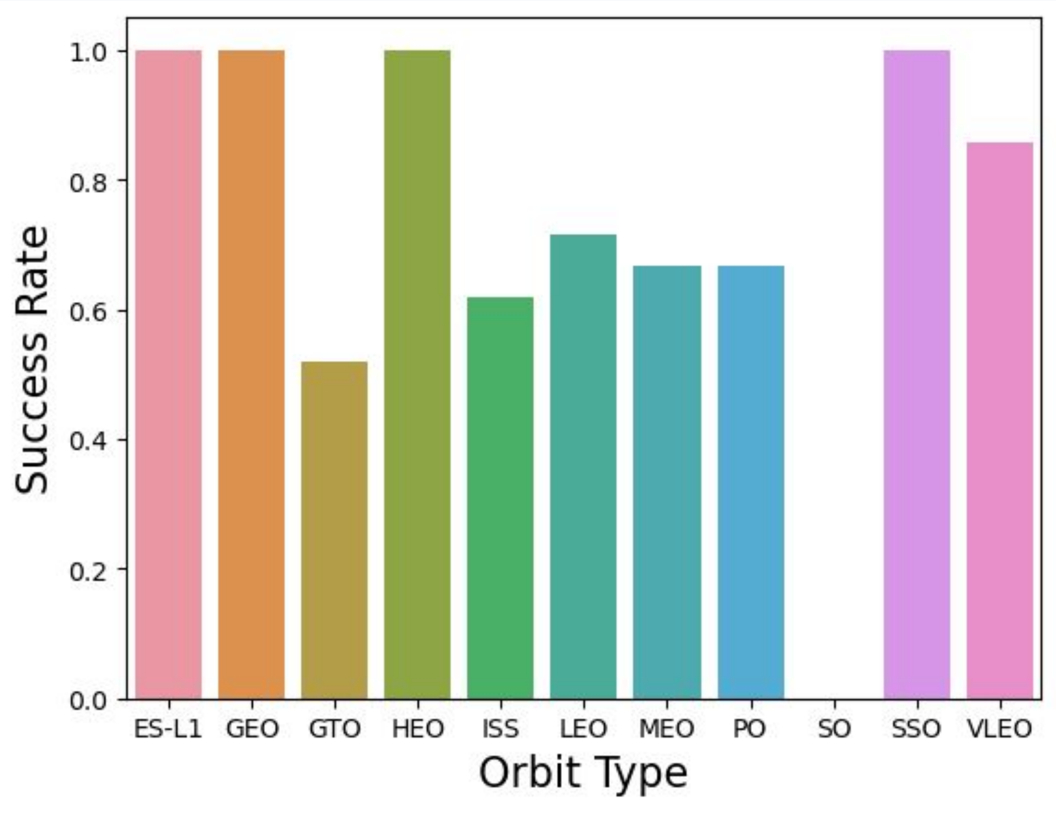


Success Rate vs. Orbit Type

The bar plot of Success Rate vs. Orbit Type indicates that:

- ES-L1, GEO, HEO and SSO have a perfect success rate.
- SO has a zero success rate.
- The remaining orbit types have a moderate to high success rate.

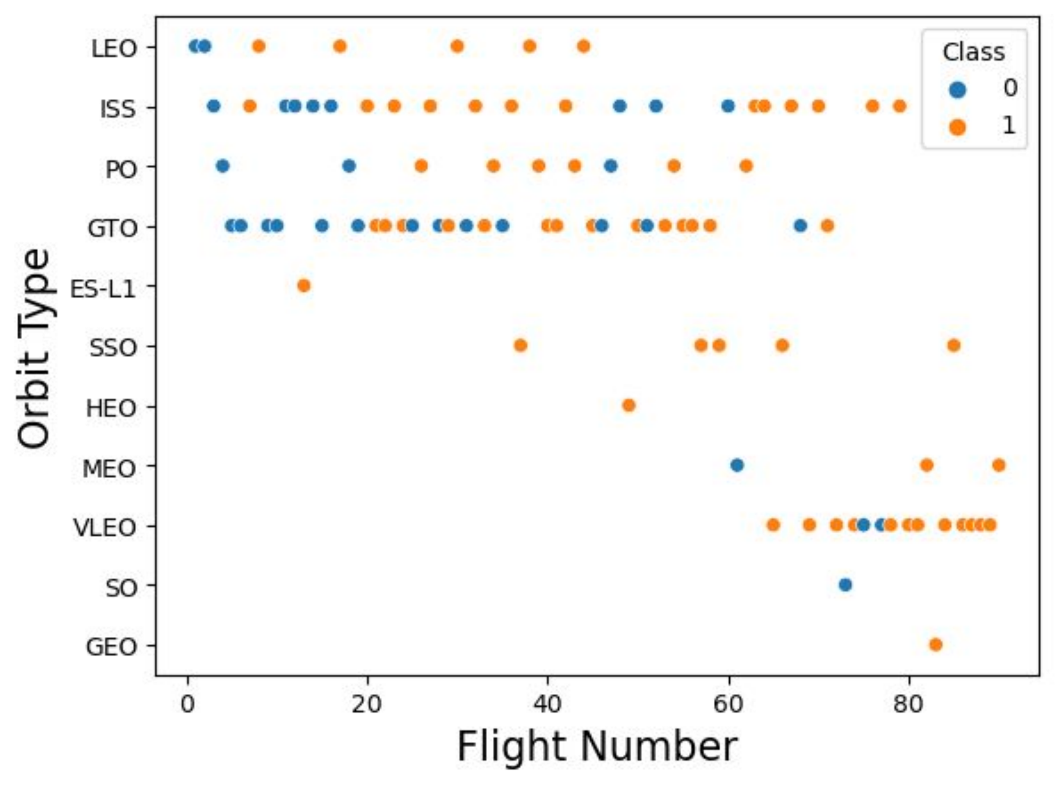
Note: the orbit categories, abbreviated here, are detailed in the Appendices.



Flight Number vs. Orbit Type

The scatter plot of Flight number vs. Orbit Type indicates that:

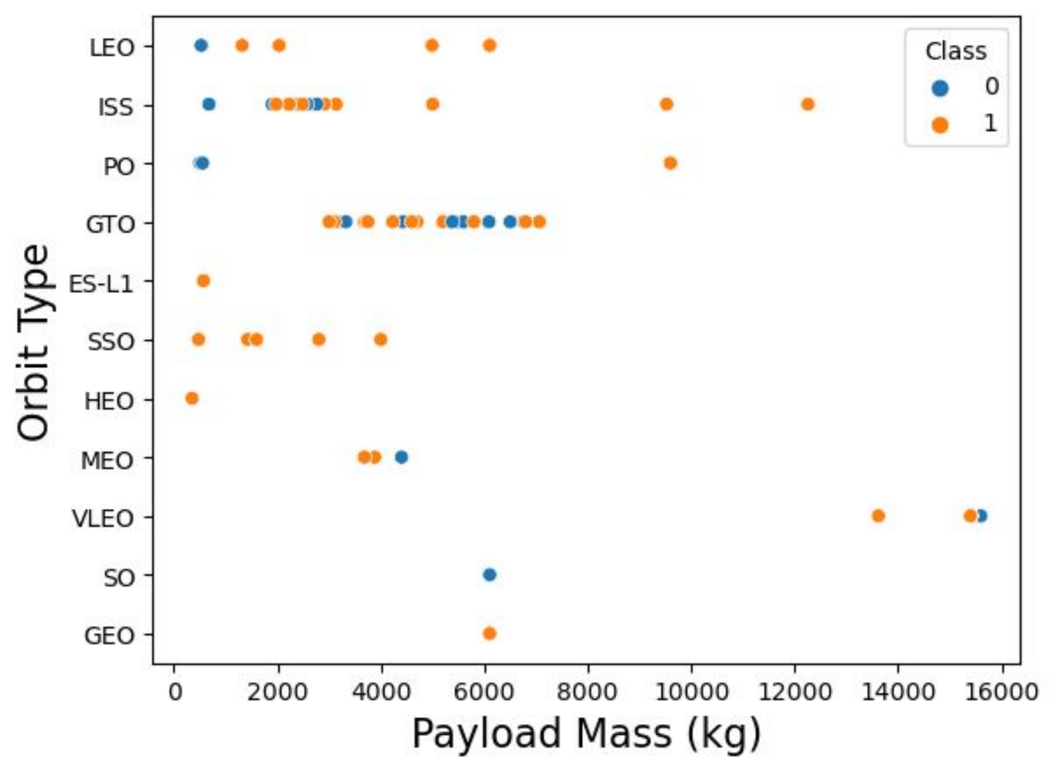
- ES-L1, SSO, HEO, MEO, SO and GEO have very few launches.
- LEO has an improved success rate with increasing flight number.
- It's not clear for some other orbit types, such as ISS and GTO, if success rate improves as flight number increases.



Payload vs. Orbit Type

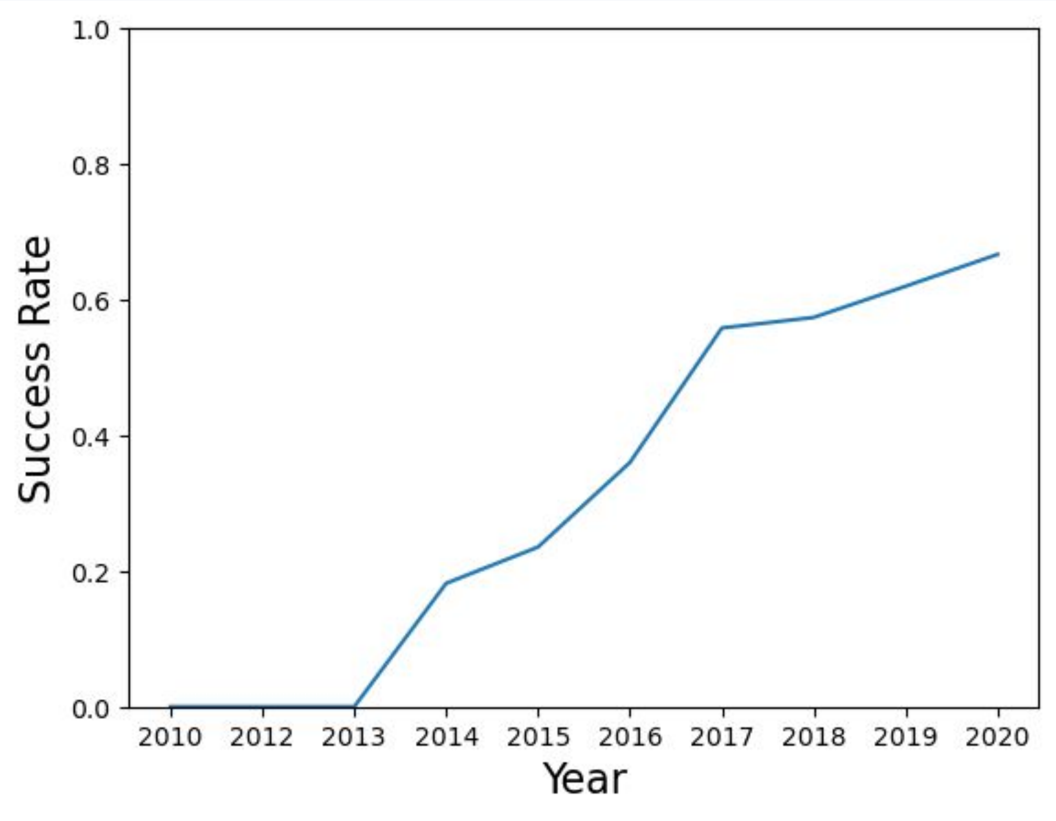
The scatter point of Payload Mass vs. Orbit Type indicates that:

- Success rate for PO, LEO and ISS increases with payload mass.
- It is not clear for GTO if success rate increases or not with payload mass.



Launch Success Yearly Trend

The line plot of yearly average success rate indicates that the success rate continually increases from 2013 to 2020.



All Launch Site Names

There are four unique launch sites used by SpaceX:

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

The following SQL query identifies the distinct names found in the launch sites column:

```
SELECT DISTINCT Launch_Site FROM SPACEXTABLE;
```

Launch Site Names Begin with 'CCA'

Below is a screenshot of five records where the launch site begins with “CCA”.

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

The following SQL query finds the full records of missions where the launch site begins with “CCA”, making use of the wildcard %, and limits the number of records to five:

```
SELECT * FROM SPACEXTABLE

WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

Total Payload Mass

The total payload mass carried by boosters from NASA (CRS) is 45,596 kg. This was found using the below SQL query, which calculated the sum of payload mass for all missions where the customer was identified to be NASA (CRS):

```
SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Customer = 'NASA (CRS)';
```

Note that CRS here stands for “Commercial Resupply Services”, a contract awarded by NASA to SpaceX for missions to resupply the International Space Station.

Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2,534.67 kg. This was found using the below SQL query, which calculated the mean payload mass for all missions where the booster version was identified to be F9 v1.1:

```
SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEXTABLE
WHERE Booster_Version LIKE 'F9 v1.1%';
```

First Successful Ground Landing Date

The first successful landing outcome on a ground pad was 22 December 2015. This was found using the below SQL query, which finds the first record where the landing outcome matched that representing a successful ground pad landing:

```
SELECT Date, Landing_Outcome FROM SPACEXTABLE
WHERE Landing_Outcome LIKE 'Success (ground pad)'
LIMIT 1;
```

Successful Drone Ship Landing with Payload between 4000 and 6000 kg

Below is a list of booster names that have successfully landed on a drone ship and had a payload mass between 4,000 and 6,000 kg:

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

This was found using the following SQL query that tested multiple conditions:

```
SELECT DISTINCT Booster_Version FROM SPACEXTABLE

WHERE (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000)

AND (Landing_Outcome = 'Success (drone ship)');
```

Total Number of Successful and Failure Mission Outcomes

In the data available, there were 100 successful and 1 failure mission outcomes, where "mission outcome" refers to the payload being delivered to its intended orbit, rather than success/failure of landing the rocket first stage.

This was found using the following SQL query, which counts the number of occurrences of each possible mission outcome:

```
SELECT Mission_Outcome, COUNT(*) AS NUM FROM SPACEXTABLE
GROUP BY Mission_Outcome;
```

Boosters Carried Maximum Payload

The names of the boosters that have carried the maximum payload mass are:

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

This was found using the following SQL query, which finds the maximum payload mass and then retrieves booster names that match that payload mass:

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

2015 Launch Records

Below is a table of selected record information for missions in 2015 that had a failure landing outcome on a drone ship:

MONTH	Landing_Outcome	Booster_Version	Launch_Site
5-	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
5-	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

This information was found using the below SQL query, which tests multiple conditions, one for the year and one for the landing outcome being a failure on a drone ship:

```
SELECT SUBSTR(Date,4,2) AS Month, Landing_Outcome,
Booster_Version, Launch_Site FROM SPACEXTABLE
WHERE (SUBSTR(Date,1,4) = '2015')
AND (Landing_Outcome LIKE 'Failure (drone ship)');
```

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

Presented is a table ranking the count of landing outcomes, in descending order, between the dates 4 June 2010 and 20 March 2017. This was found using the below SQL query, which counted the occurrence of each landing outcome in the desired date range:

```
SELECT Landing_Outcome, COUNT(*) FROM SPACEXTABLE
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY Landing_Outcome
ORDER BY NUM DESC;
```

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

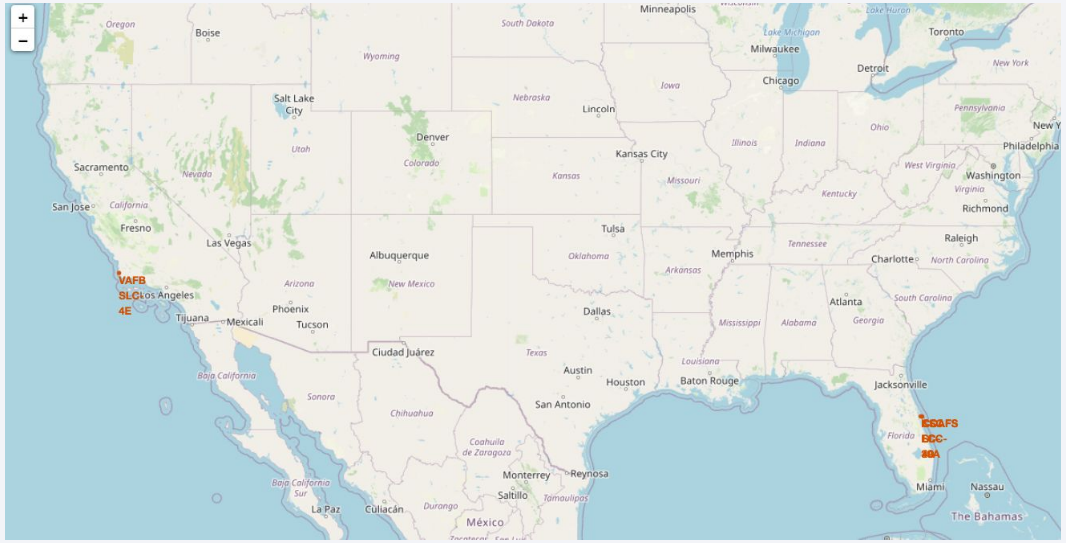
Section 3

Launch Sites Proximities Analysis

Map of SpaceX Launch Sites

As seen in the displayed map, all launch sites are close to the equator and to the coastline.

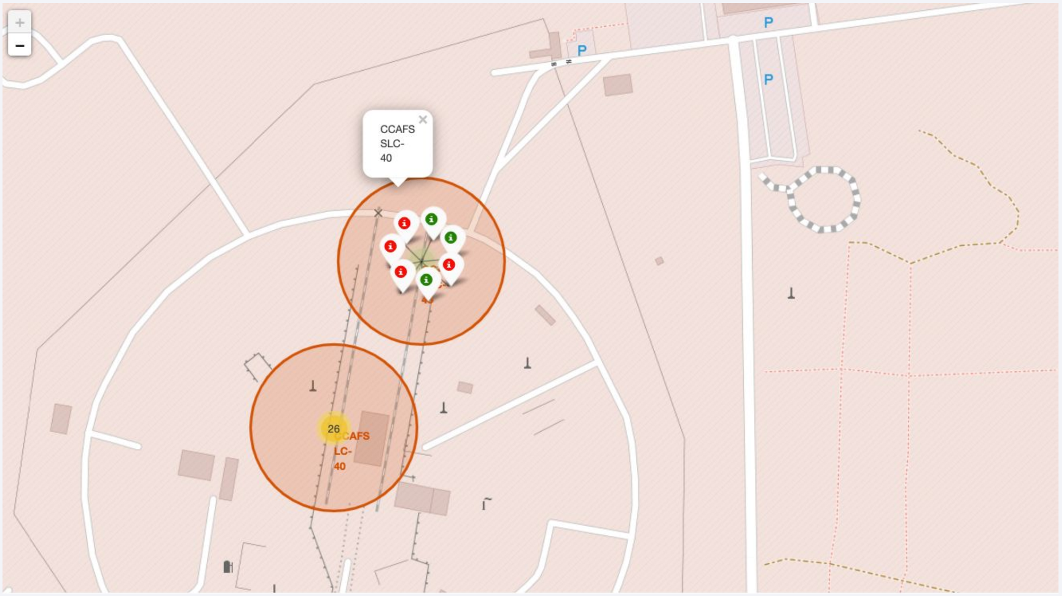
Please note that three of the four launch sites are located close together in Florida, causing their name labels to overlap.



Success/Fail Launches at Launch Sites

Shown in the displayed map are the successful (green) and failure (red) landing outcomes for all missions that launched from CCAFS SLC-40.

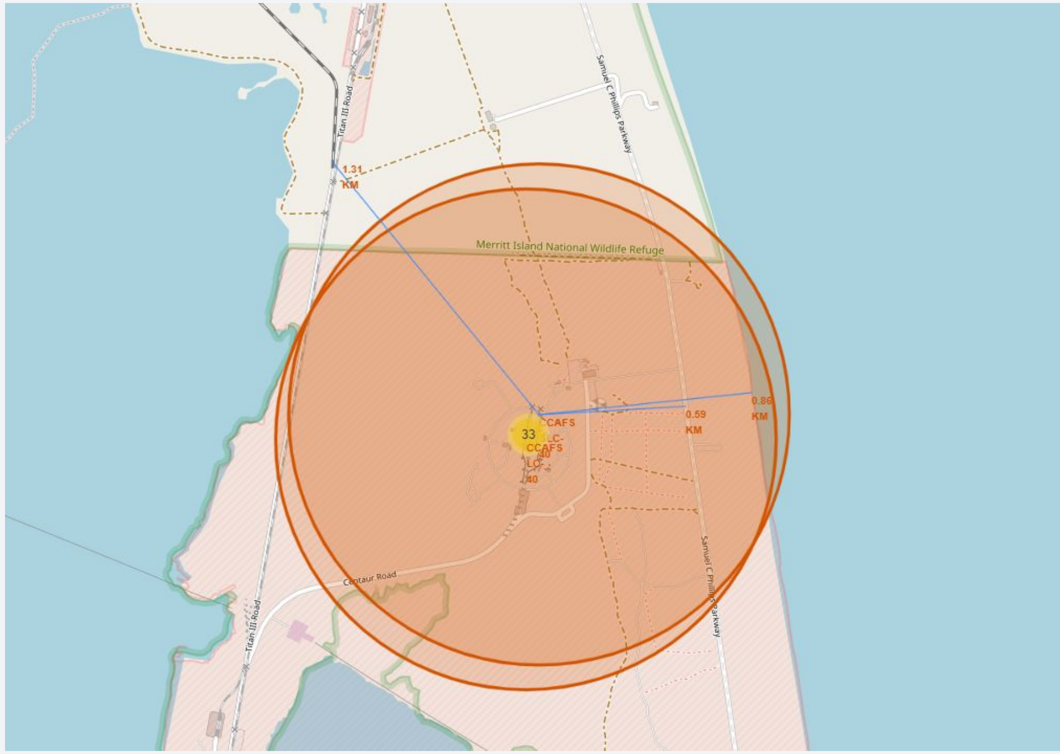
This launch site is indicated by a red transparent circle and is located very close to CCAFS LC-40. It can be seen from the yellow circle that there have been 26 launches at that site.



Proximity of Points of Interest to Launch Site

Displayed in the map are the launch sites of CCAFS SLC-40 and CCAFS LC-40, with blue lines showing the proximity of the former to nearby points of interest:

- Railway (top-left): 1.31 km
- Coastline (upper middle): 0.86 km
- Highway (lower middle): 0.59 km



Section 4

Build a Dashboard with Plotly Dash

Launch Success for All Launch Sites

The pie chart shown indicates that the number of successful launches at each site, as a percentage of the total number of successful launches, are:

- KSC LC-39A: 41.7%
- CCAFS LC-40: 29.2%
- VAFB SLC-4E: 16.7%
- CCAFS SLC-40: 12.5%

Successful vs Failed Launches at KSC LC-39A

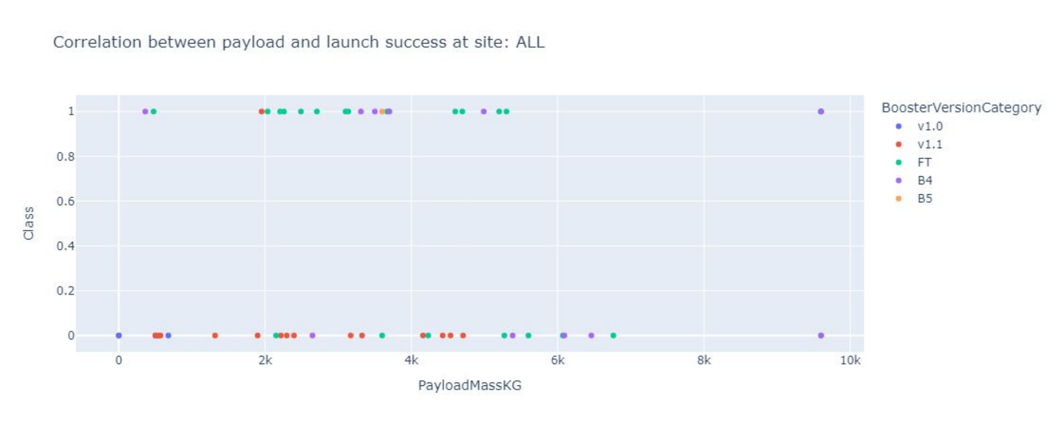
The pie chart shown indicates the relative percentage of successful vs failed launches from the site that has the highest number of successful launches.

It is clear that a majority of launches from this site are successful, at 76.9%.

<Dashboard Screenshot 3>

The top scatter plot shows the payload mass vs launch outcome (0=fail, 1=success) for all sites, coloured by booster version.

The bottom scatter plot shows that a payload mass range of 0 to 5500 kg at site KSC LC-39A has a perfect success rate across 10 launches.



Section 5

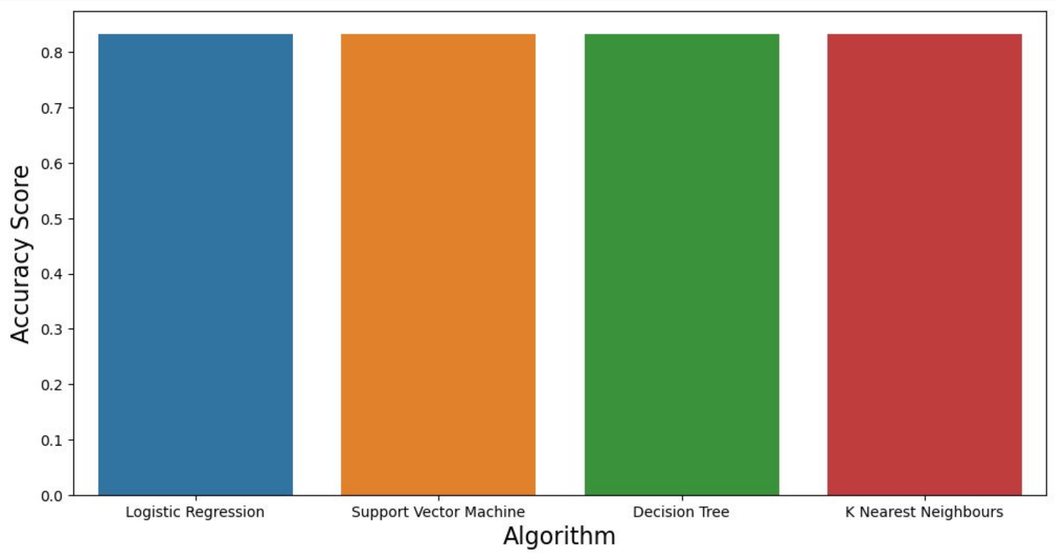
Predictive Analysis (Classification)

Classification Accuracy

Shown is a bar chart of the accuracy for each algorithm considered.

Accuracy here is simply a measure of of correctly predicted outcomes in the test data set.

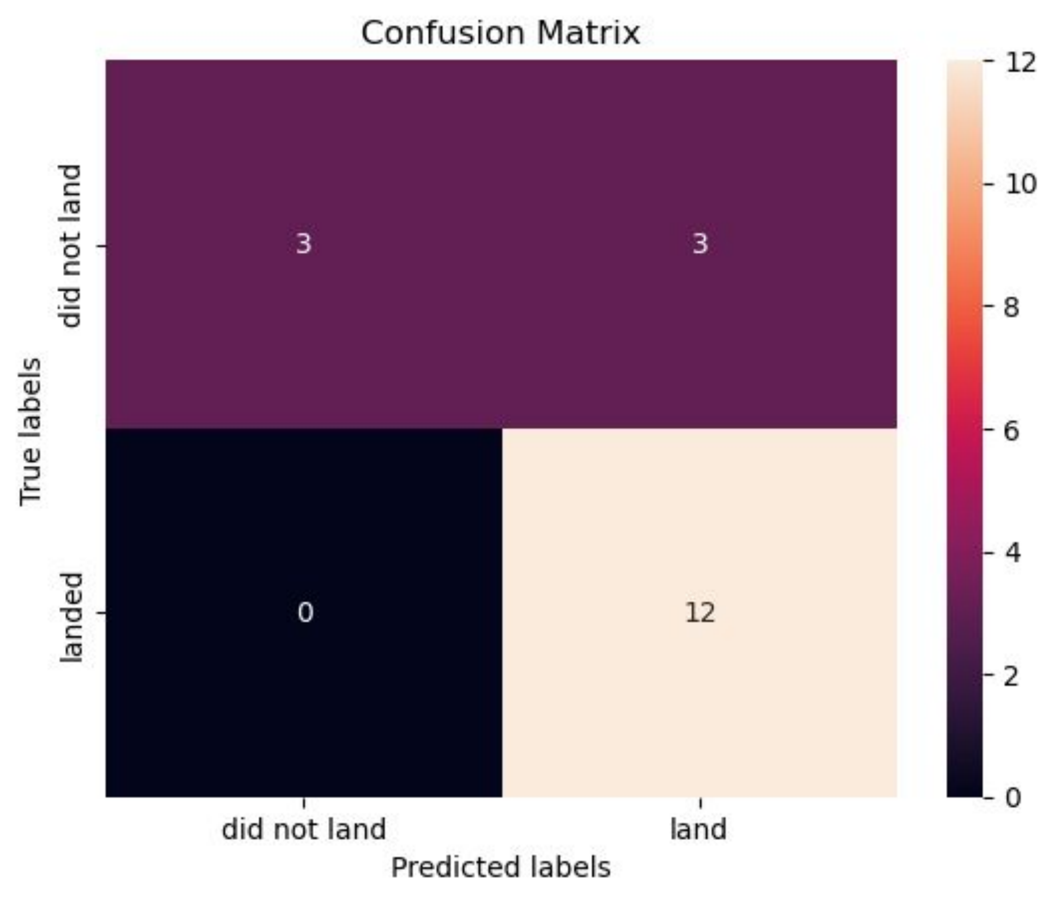
Interestingly, all algorithms return the same accuracy of 0.83 i.e. 83% of outcomes were correctly predicted.



Confusion Matrix

Shown is the confusion matrix for the Support Vector Machine algorithm.

The diagonal values show the correctly predicted outcomes. All successful landings were correctly predicted, but only half of failure landings were predicted correctly i.e. the model has a high rate of false negatives.



Conclusions

Sufficient data exists through the SpaceX API to understand, model and predict the outcome of rocket first stage landing success.

There are a number of features that appear to impact whether a landing will be successful, including payload mass and orbit type. The success rate has gradually increased over time, up to around 60-70% in recent years.

It is possible to predict, with a reasonable level of accuracy, the landing outcome based on the available features using machine learning algorithms. Successful landings in particular can be predicted very accurately.

Appendices: Relevant URLs

GitHub Repository containing all Jupyter Notebooks and associated data and files.

Data collection URLs:

- SpaceX API: Launch data, Launch sites, Payload data, Booster version, Core data
- Web scraping

Appendices: Orbit Types

There are 11 orbit categories:

- LEO: Low Earth Orbit
- VLEO: Very Low Earth Orbits
- GTO: Geosynchronous Orbit
- SSO/SO: Sun-synchronous Orbit
- ES-L1: Lagrange-point Orbit
- HEO: Highly Elliptical Orbit
- ISS: International Space Station
- MEO: Intermediate Circular Orbit
- HEO: High Geocentric Orbit
- GEO: Circular Geocentric Orbit
- PO: Polar Orbit

Thank you!