Winning Space Race with Data Science

Presented by Jason Lau

24th May, 2022

Outline

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

Executive Summary

Data was collected from SpaceX API and scraped from Wikipedia, and further cleaned and wrangled. After that exploratory data analysis was carried out and crucial features were selected for classification modeling.

For exploratory data analysis, it was, firstly, found that there are four unique launch sites for Falcon 9 launch where KSC LC-39A have the greatest success launch, and they are all located near the coastline and away from cities owing to safety consideration. Secondly, ES-L1, SSO, HEO and GEO orbits have the greatest success launch. Thirdly, the success rate increases from 2013 to 2020. Furthermore, vital features including flight number, payload mass, orbit, launch site, flights, grid fins, reused, legs, landing pad, block, reused count and serial were selected. It was, then, moved on to the classification modeling. All the models trained including logistic regression, support vector machine, decision tree and k nearest neighbors for the Falcon 9 successful landing prediction have the highest accuracy of 0.833.

Introduction

Background

As the development of technology has become more advanced, space travel becomes more accessible and affordable for everyone.

Competitors

- SpaceX
- Virgin Galactic
- Rocket Lab
- Blue Origin

Reasons for SpaceX is the strongest competitor

- First stage of their rocket launch (Falcon 9) is reused
  - Determining stage of a rocket launch (vs. second stage)
- SpaceX spend an inexpensive price, 62 million dollars
- Other competitors in general, 165 million dollars

Project aim

To predict the first stage of Falcon 9 from SpaceX lands successfully

- The cost of a rocket launch can then be determined

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collect raw Falcon 9 historical launch record by requesting for the SpaceX API and web scraping from a Wikipedia page
- Perform data wrangling
  - Replace null values of columns (containing numeric continuous values) to the mean of the column
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Build logistic regression, support vector machine, decision tree and k nearest neighbors models
  - Tune and optimize the hyperparameters using grid search
  - Evaluate and choose the model with the highest test accuracy

Data Collection – SpaceX API

SpaceX API

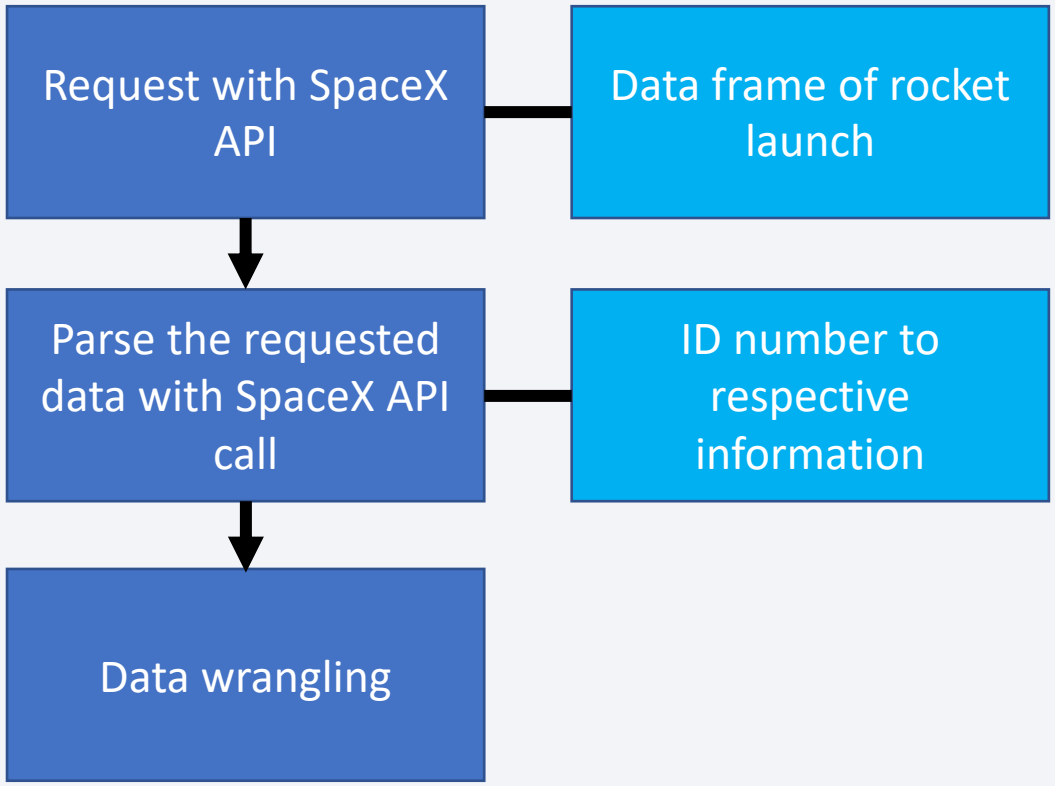
https://api.spacexdata.com/v4/

- Request the rocket launch data from the URL
- Parse the requested data to turn the identification numbers in the column to respective useful information

Data wrangling

- Include Falcon 9 data only
- Replace null values

My work: Data Collection – SpaceX API

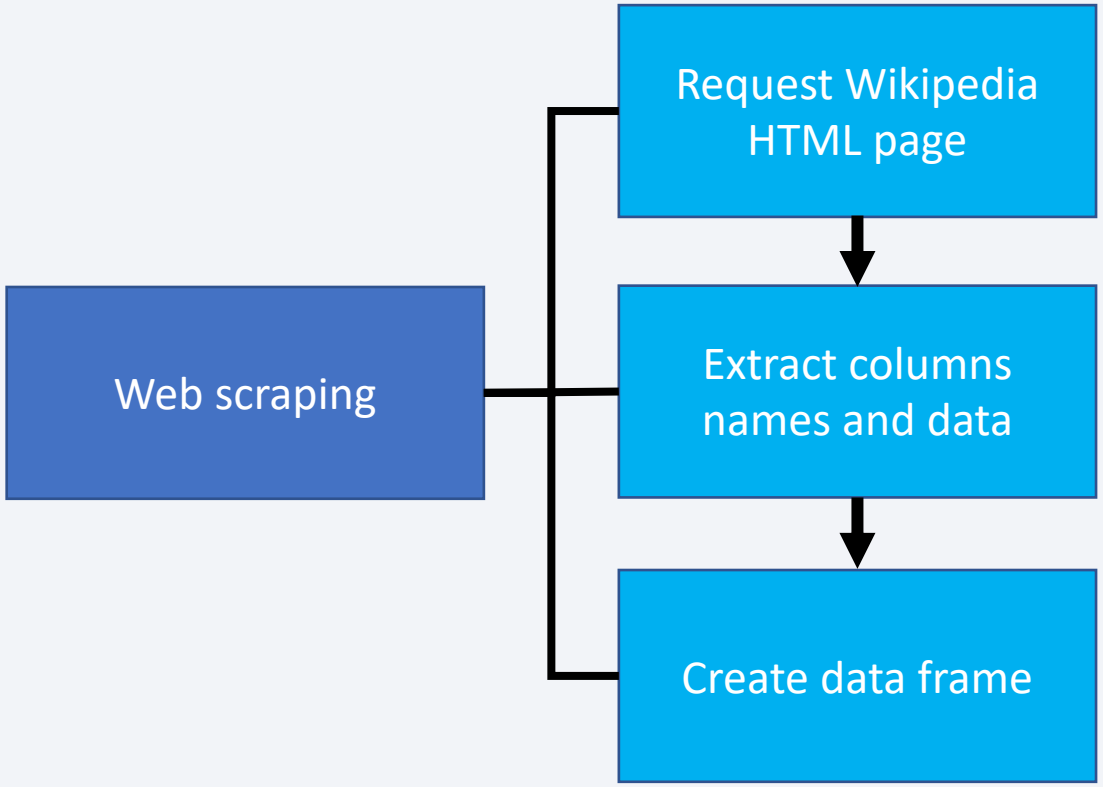


Data Collection - Scraping

Web scraping

https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

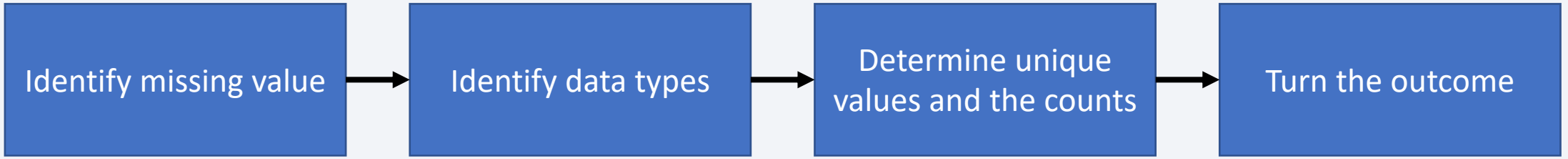
- Request the Falcon 9 Launch Wikipedia HTML page from the URL
- Extract column names and data from the HTML table header
- Create a data frame by parsing the HTML table



My work: Data Collection – Scraping

Data Wrangling

- Identify missing values in each attribute
- Identify data type of each attribute
  - Numeric/categorical
- Determine the unique values and their counts of each attribute
- Turn the categorical outcomes into Boolean value
  - Successful (1)/unsuccessful (0)



My work: Data Wrangling

EDA with Data Visualization

To observe if there is any relationship between different combinations of attributes that influence success rate

- Launch site against flight number
- Launch site against payload
- Success rate against orbit type
- Orbit type against flight number
- Orbit type against payload mass
- Launch success yearly trend

To select important attributes for further machine learning

My work: EDA with Data Visualization

EDA with SQL

- Unique values of launch site
- Total number of success and failure outcome
- Unique value and count of landing outcome
- Maximum and average payload mass given booster version of the rocket

My work: EDA with SQL

Build an Interactive Map with Folium

Map object created

- Markers and circles: all launch sites
- Marker: success and failure outcomes
- Lines: distance between a launch site and places
  - Highways
  - Coastlines
  - Railways
  - Cities

To investigate the common features from locations of each launch site and the distance between different places

My work: Interactive Map with Folium

Build a Dashboard with Plotly Dash

Plots to be created

- Pie chart: success rate of rocket launch at different launch sites
- Scatter chart: success rate of rocket launch given a range of payload mass

To investigate and compare the success rate at different launch sites and given different ranges of payload

My work: Dashboard with Plotly Dash

Predictive Analysis (Classification)

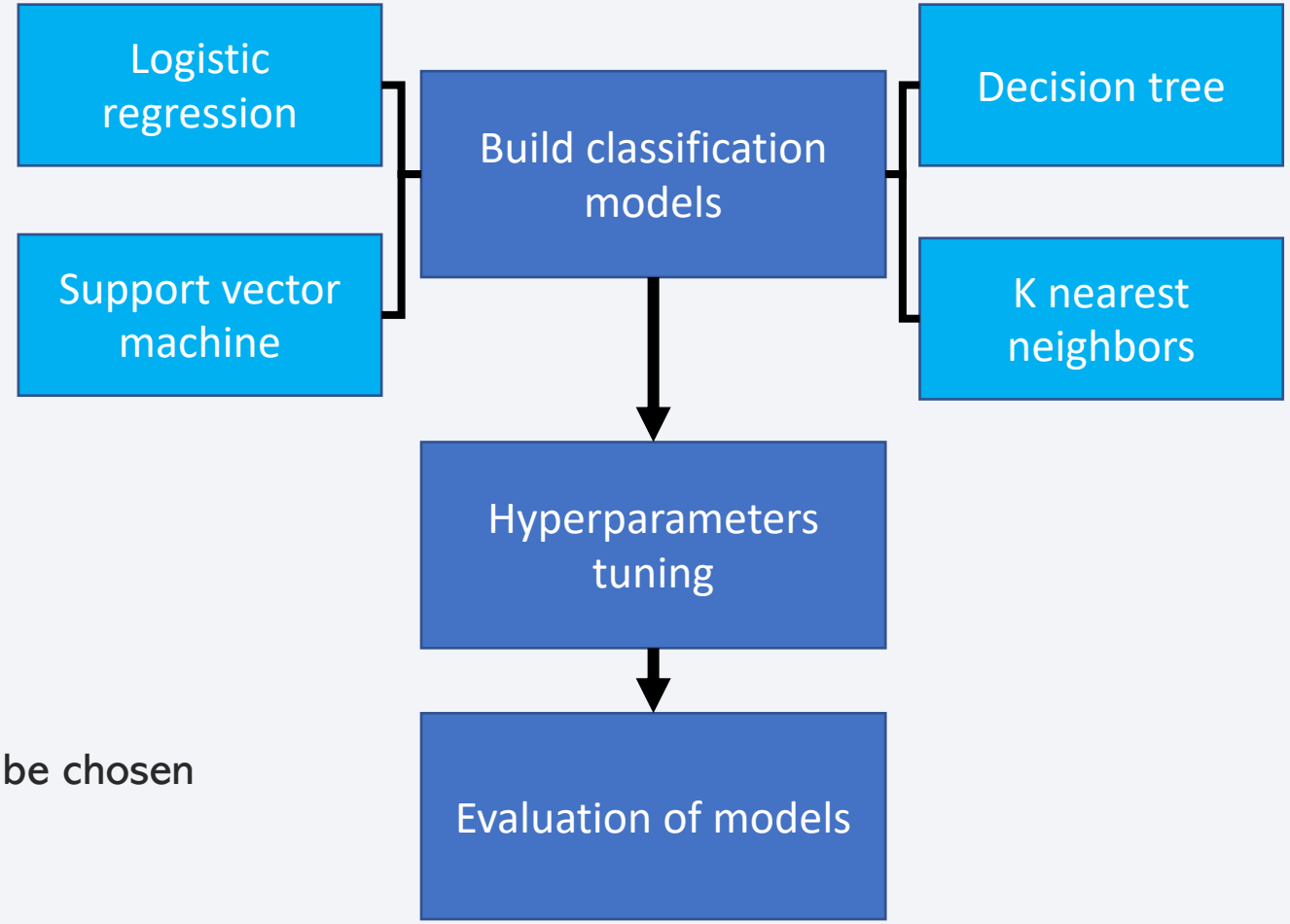
Classification models to be built
- Logistic regression
- Support vector machine
- Decision tree
- K nearest neighbors

Hyperparameters tuning
- Grid search

Evaluation of models
- Accuracy score
- Confusion matrix

The model with the highest test accuracy score will be chosen

My work: Predictive analysis

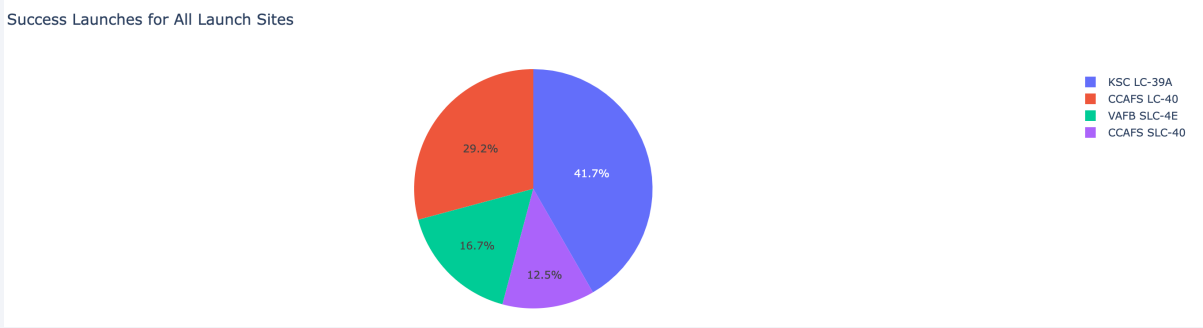


Results

Exploratory data analysis

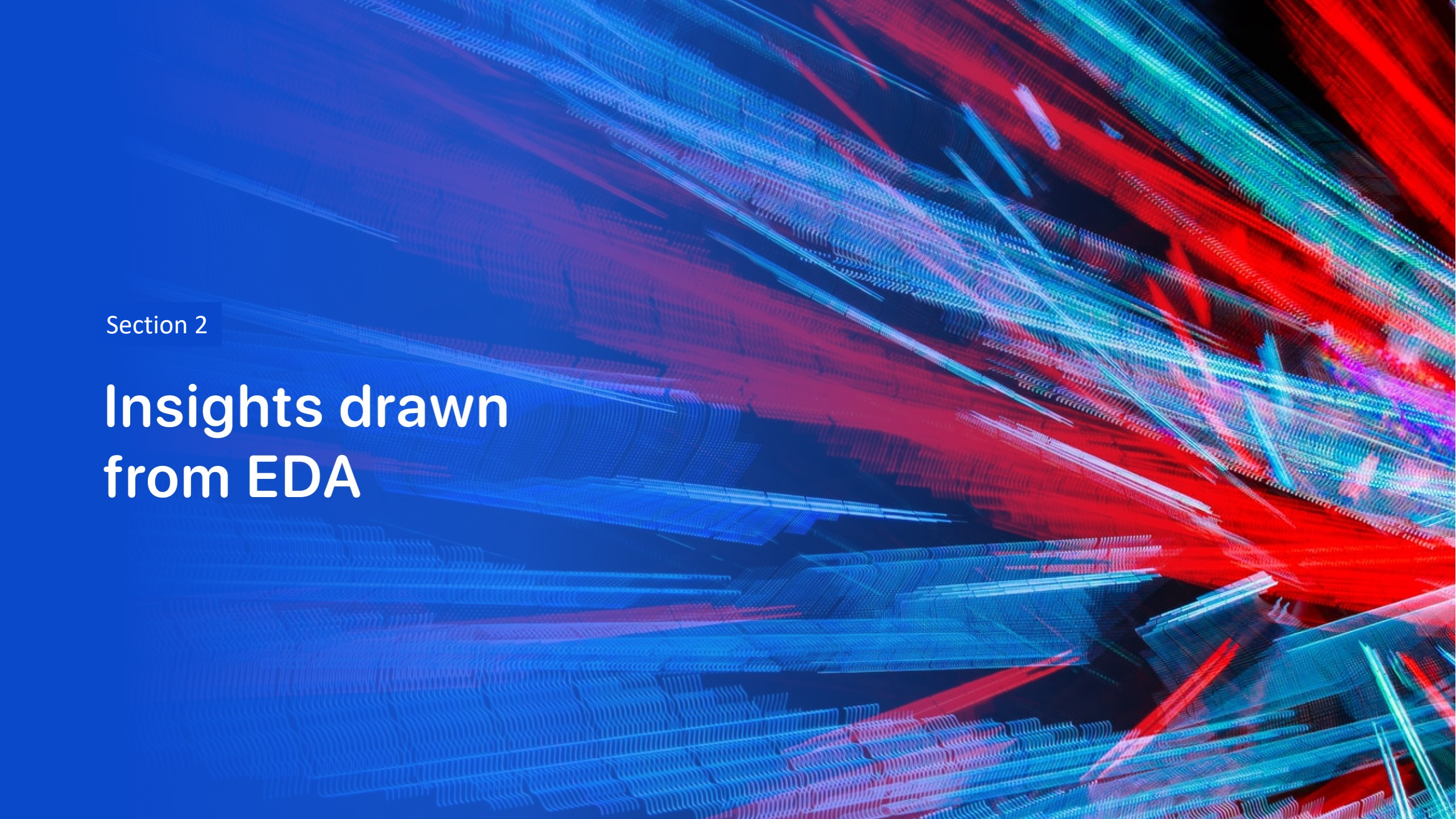
- Important features were identified for machine learning

Interactive analytics



Predictive analysis

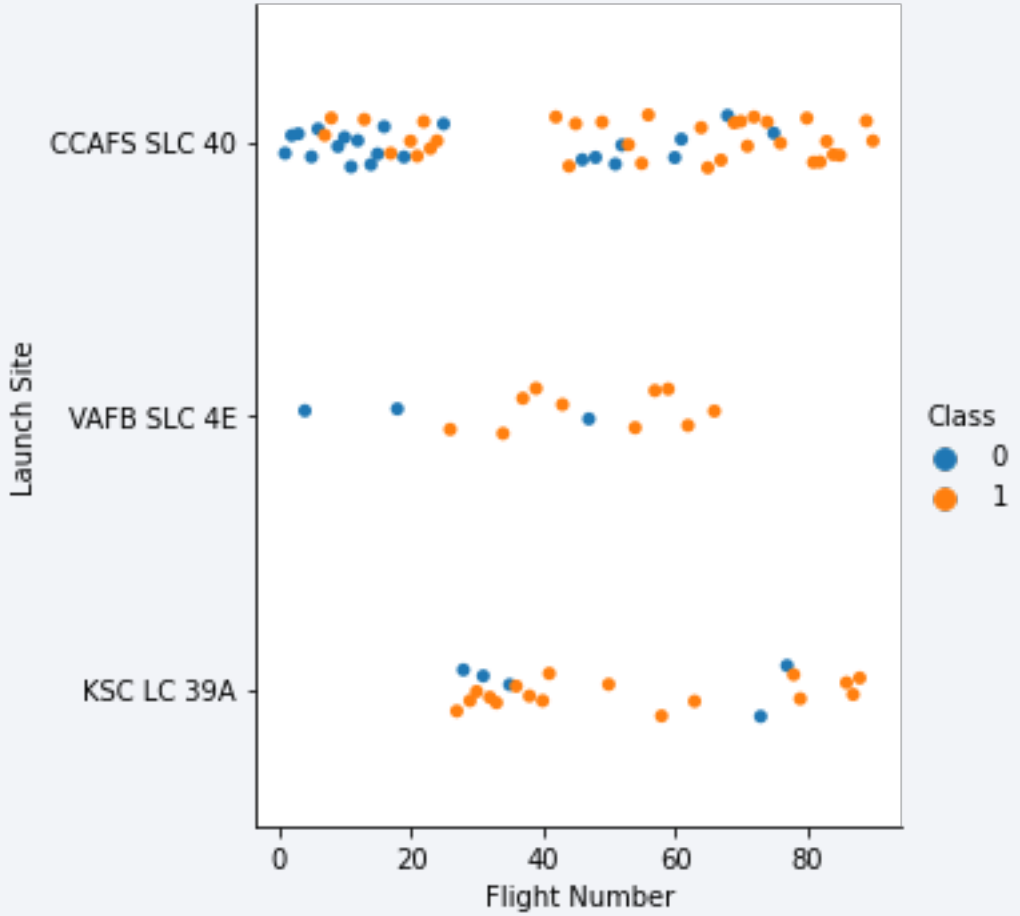
- All models could be deployed as they all gave accuracy of 0.83


Section 2
Insights drawn from EDA

Flight Number vs. Launch Site

When the flight number increases, the success rate increases more obviously at VAFB SCL 4E than those of the other two launch sites

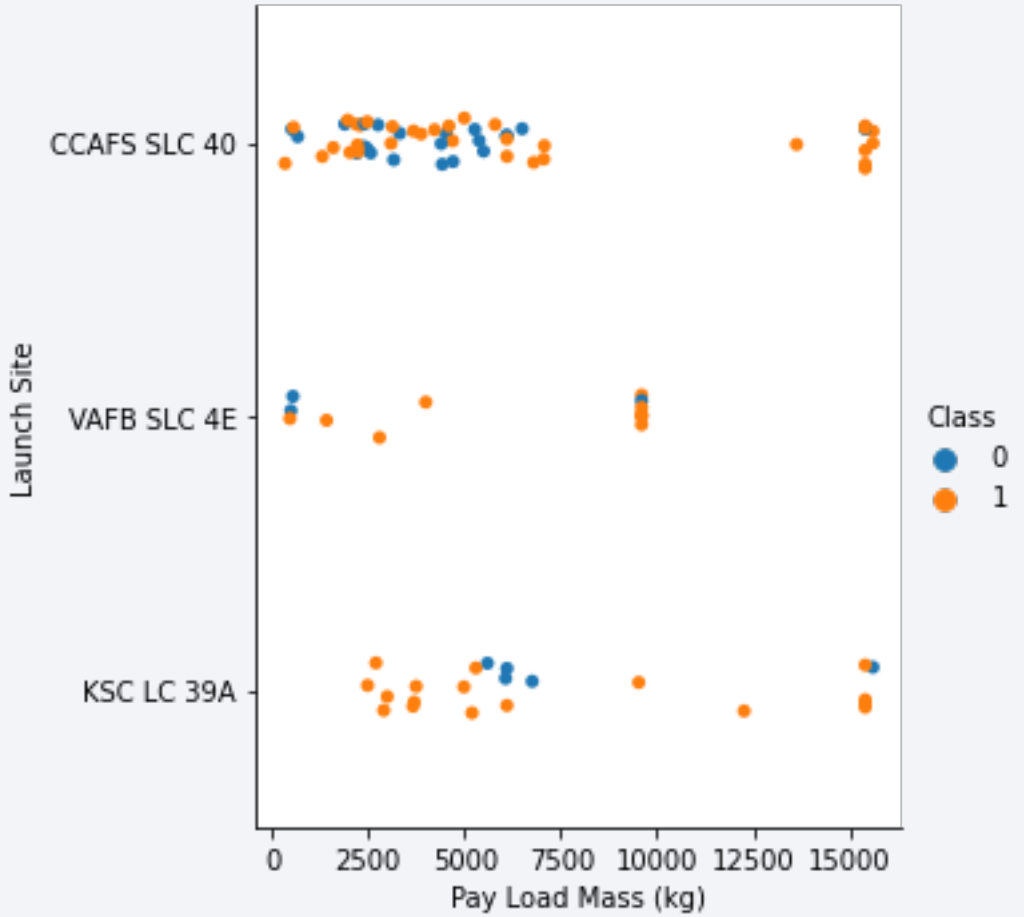
However, there is no significant relationship between flight number and success rate



Payload vs. Launch Site

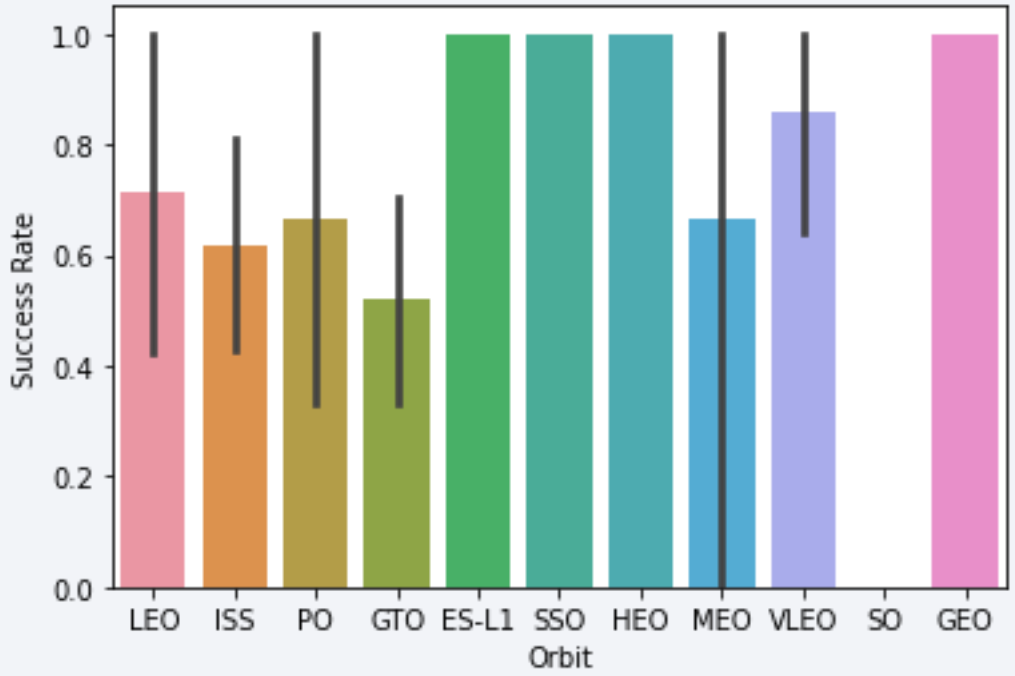
At VAFB SLC 4E, there is no payload mass greater than 10,000 kg

At CCAFS SLC 40 and KSC LC 39A, the proportion of success rate is higher when the payload mass is greater than 10,000 kg compared to that is smaller than 10,000 kg



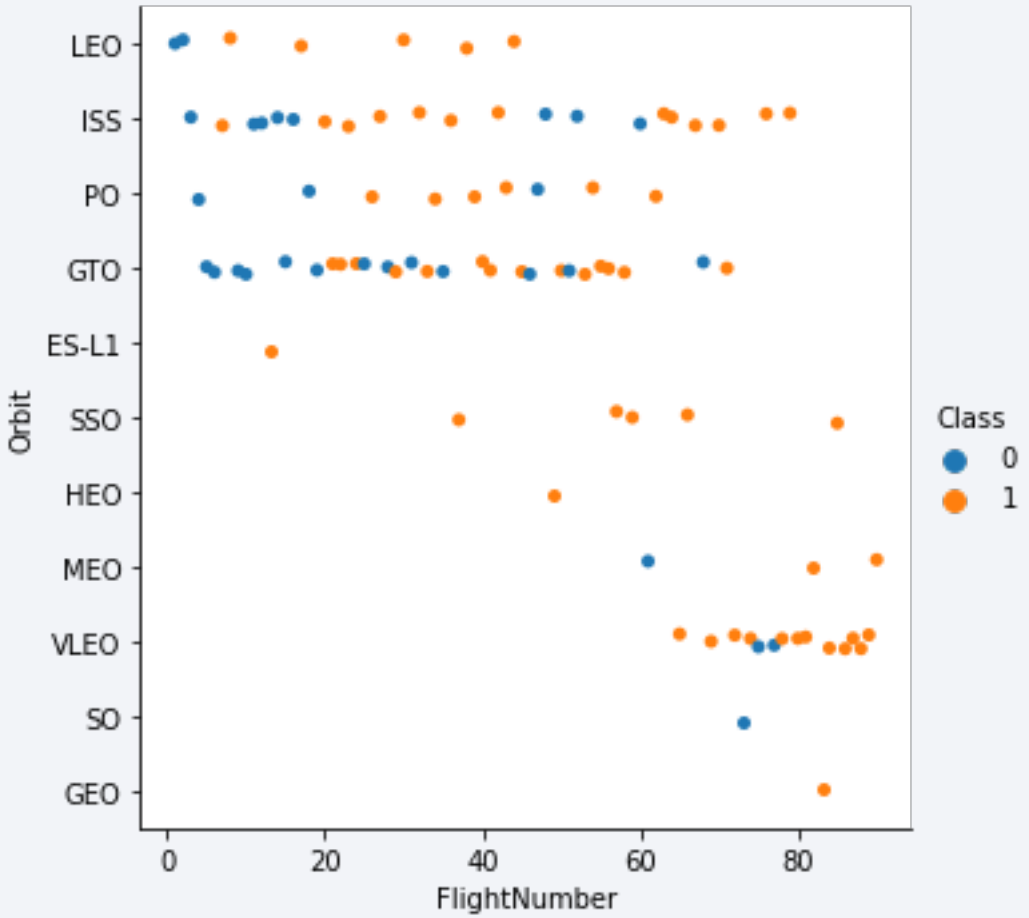
Success Rate vs. Orbit Type

ES-L1, SSO, HEO and GEO have the highest success rate



Flight Number vs. Orbit Type

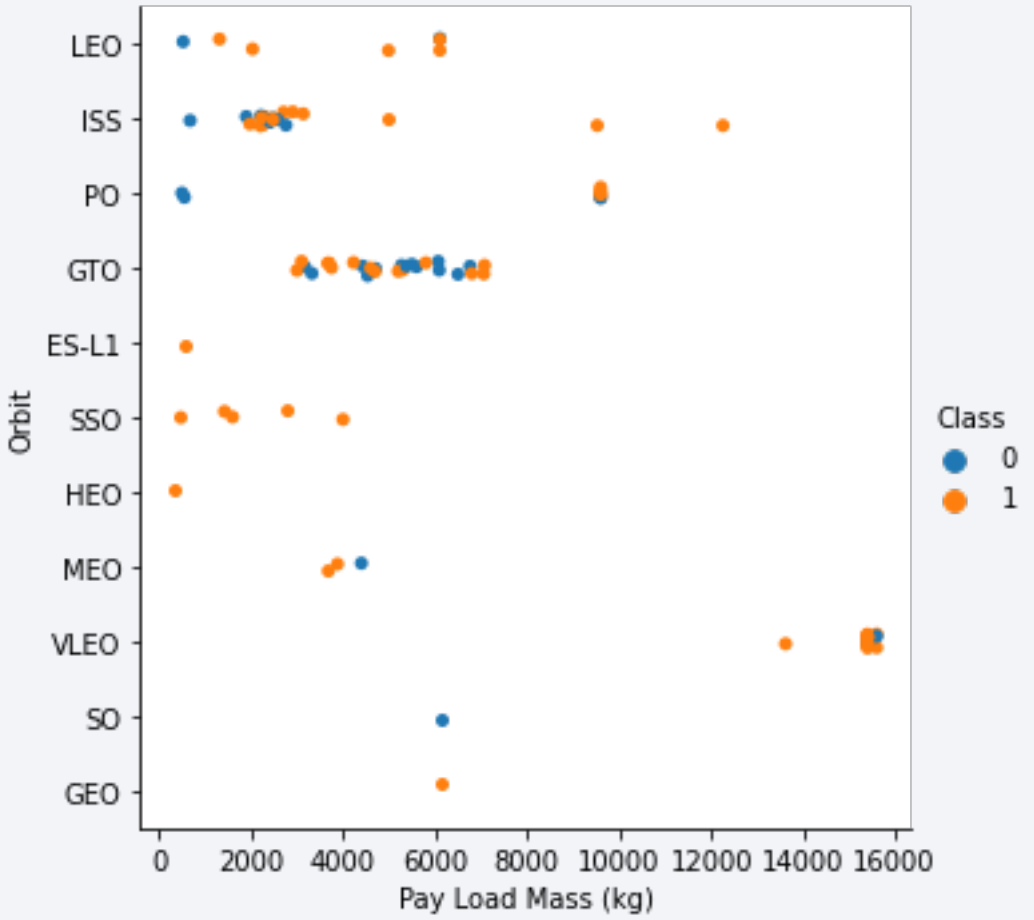
For orbit LEO, the flight number relates to the success rate while there is weak relationship between success rate and orbits (ISS, PO, GTO and VLEO)



Payload vs. Orbit Type

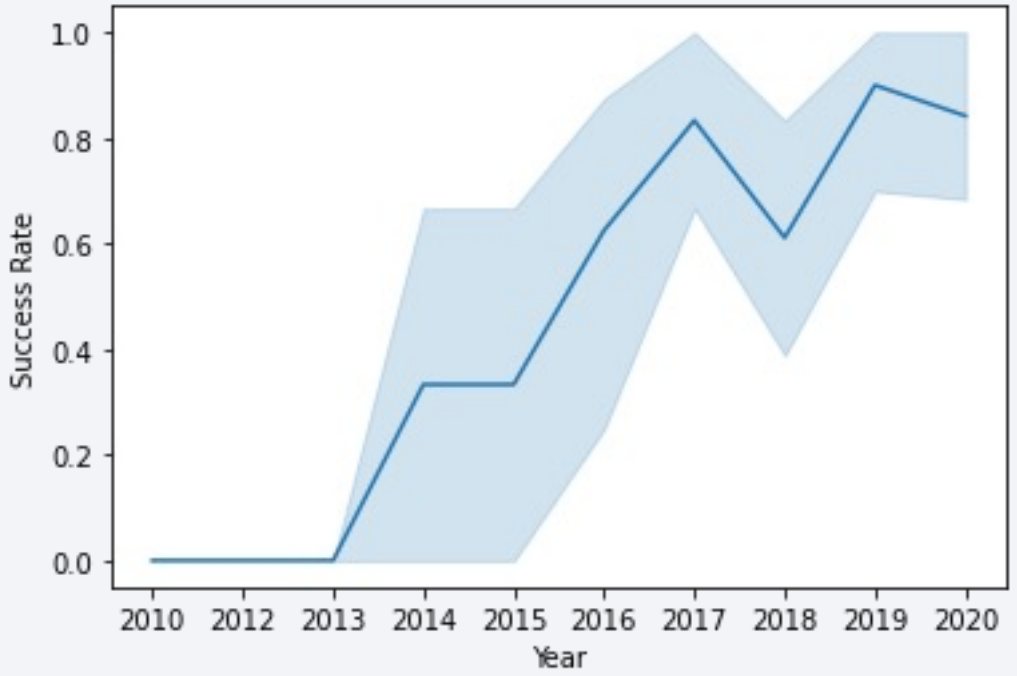
For orbits LEO and ISS, the success rate is higher when the payload mass increases

However, there is no significant relationship between the payload mass and the success rate for orbit GTO



Launch Success Yearly Trend

The success rate increases from 2013 to 2020

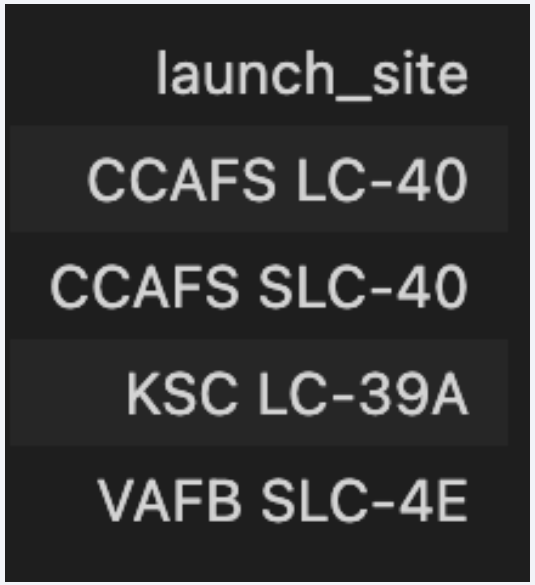


All Launch Site Names

Unique launch sites

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

Understand the the number and unique launch site for falcon 9 launch



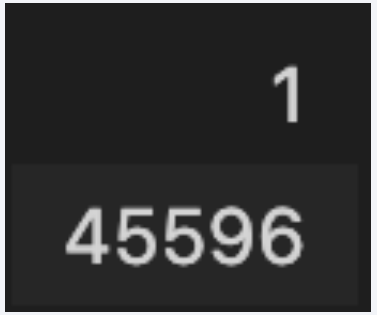
Launch Site Names Begin with 'CCA'

To have a glance of the information given a launch site that begins with ‘CCA’

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

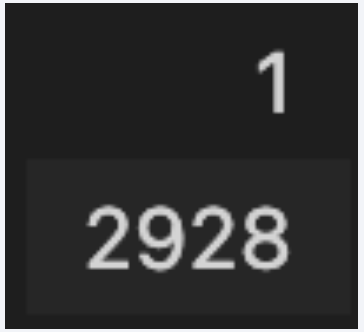
Total Payload Mass

The total payload carried by boosters from NASA is 45,596 kg



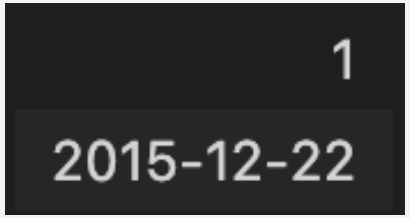
Average Payload Mass by F9 v1.1

The average payload mass carried by booster version F9 v1.1 is 2,928 kg



First Successful Ground Landing Date

The first successful landing outcome on ground pad is 2015-12-22



Successful Drone Ship Landing with Payload between 4000 and 6000

The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are

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

booster_version	landing__outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

Total Number of Successful and Failed Mission Outcomes

The total number of successful mission outcome is 100 with one that the payload status remains unclear while that of the failure outcomes is 1

mission_outcome	2
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

The names of the booster which 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 B1049.5
- F9 B5 B1060.2
- F9 B5 B1058.3
- F9 B5 B1051.6
- F9 B5 B1060.3
- F9 B5 B1049.7

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

2015 Launch Records

The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015 is F9 v1.1 B1012 and F9 v1.1 B1015 at CCAFS LC-40

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

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

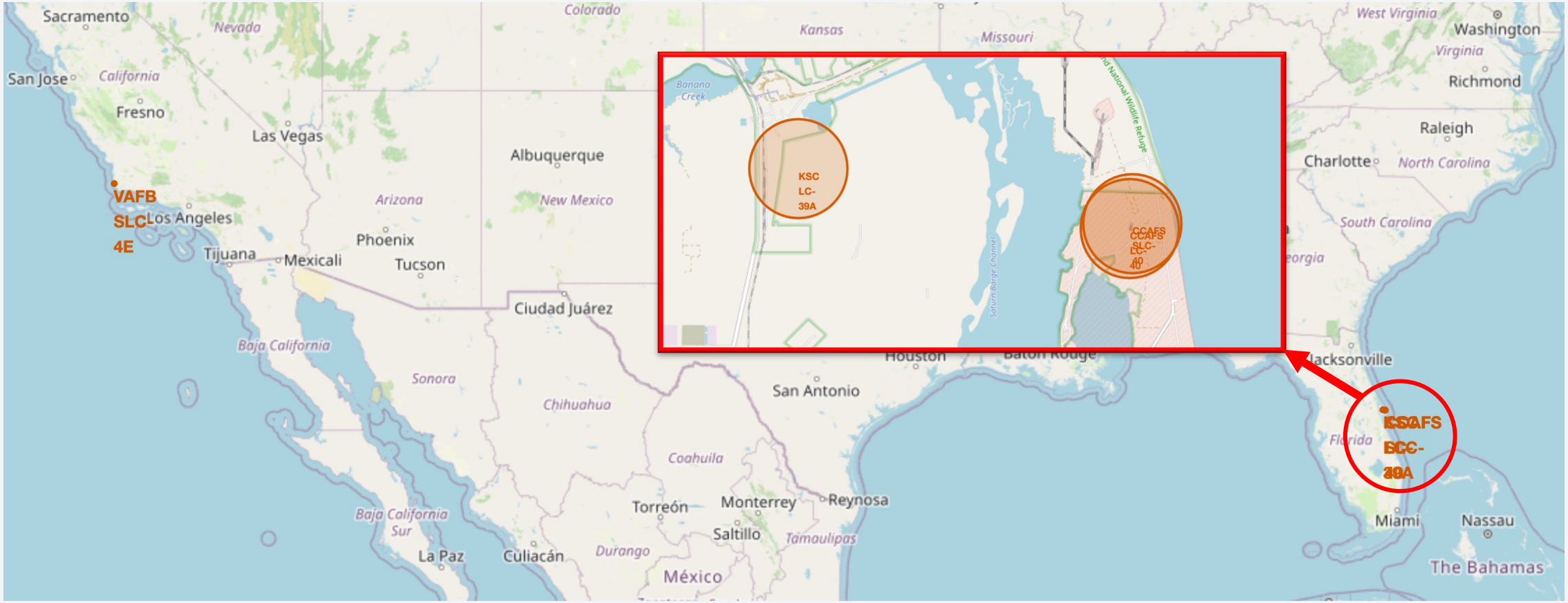
The landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order are

- No attempt
- Failure (drone ship) and success (drone ship)
- Controlled (ocean) and success (ground pad)
- Failure (parachute) and uncontrolled (ocean)
- Precluded (drone ship)

landing__outcome	2
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


Section 3
Launch Sites
Proximities Analysis

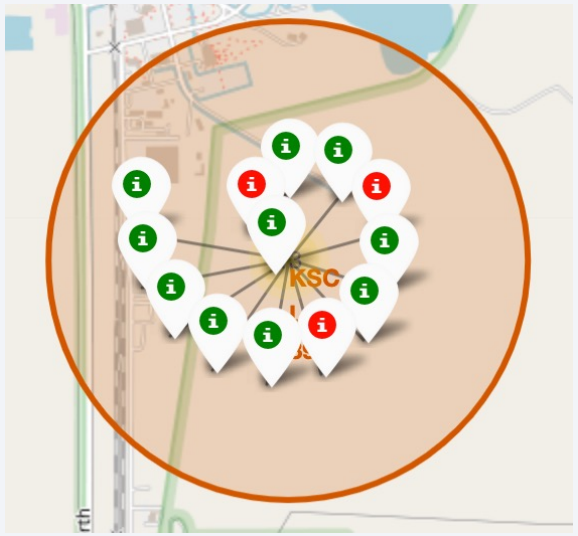
Folium Map: all launch sites



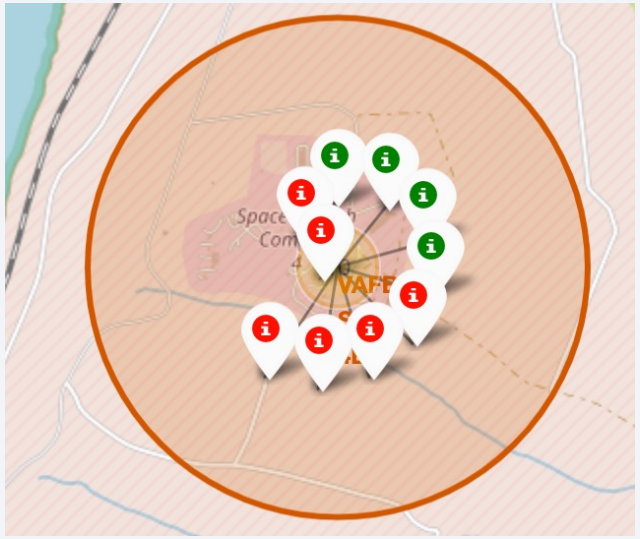
All the launch sites are located near the coastlines

Folium Map: launch outcomes at launch sites

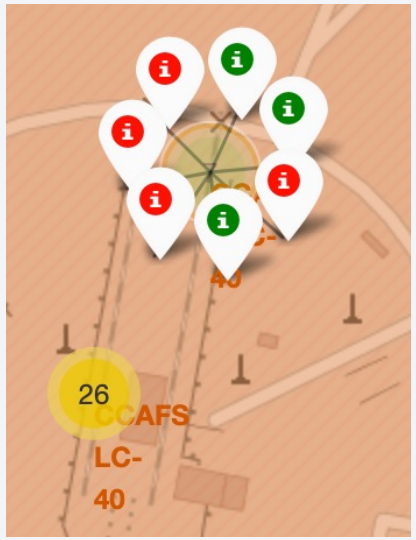
success rate at KSC LC-39A is the highest compared to that of VAFB SLC-4E, CCAFS SLC 40 and CCAFS LC 40



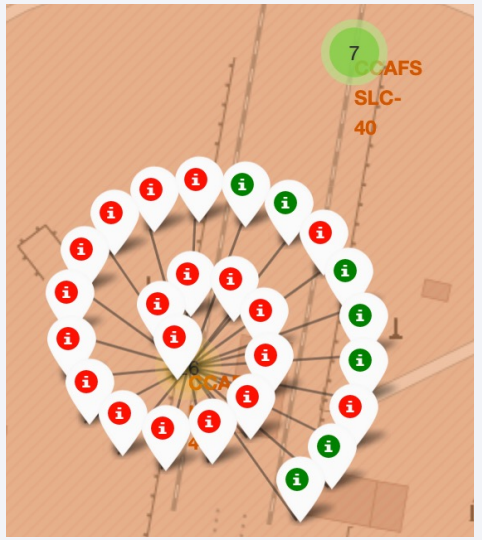
KSC LC-39A



VAFB SLC-4E



CCAFS SLC 40

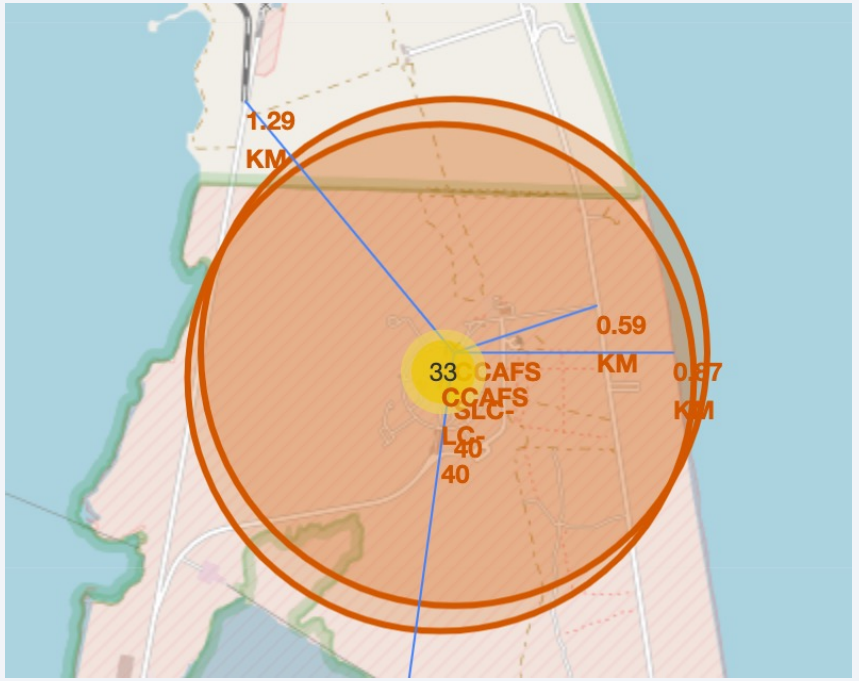


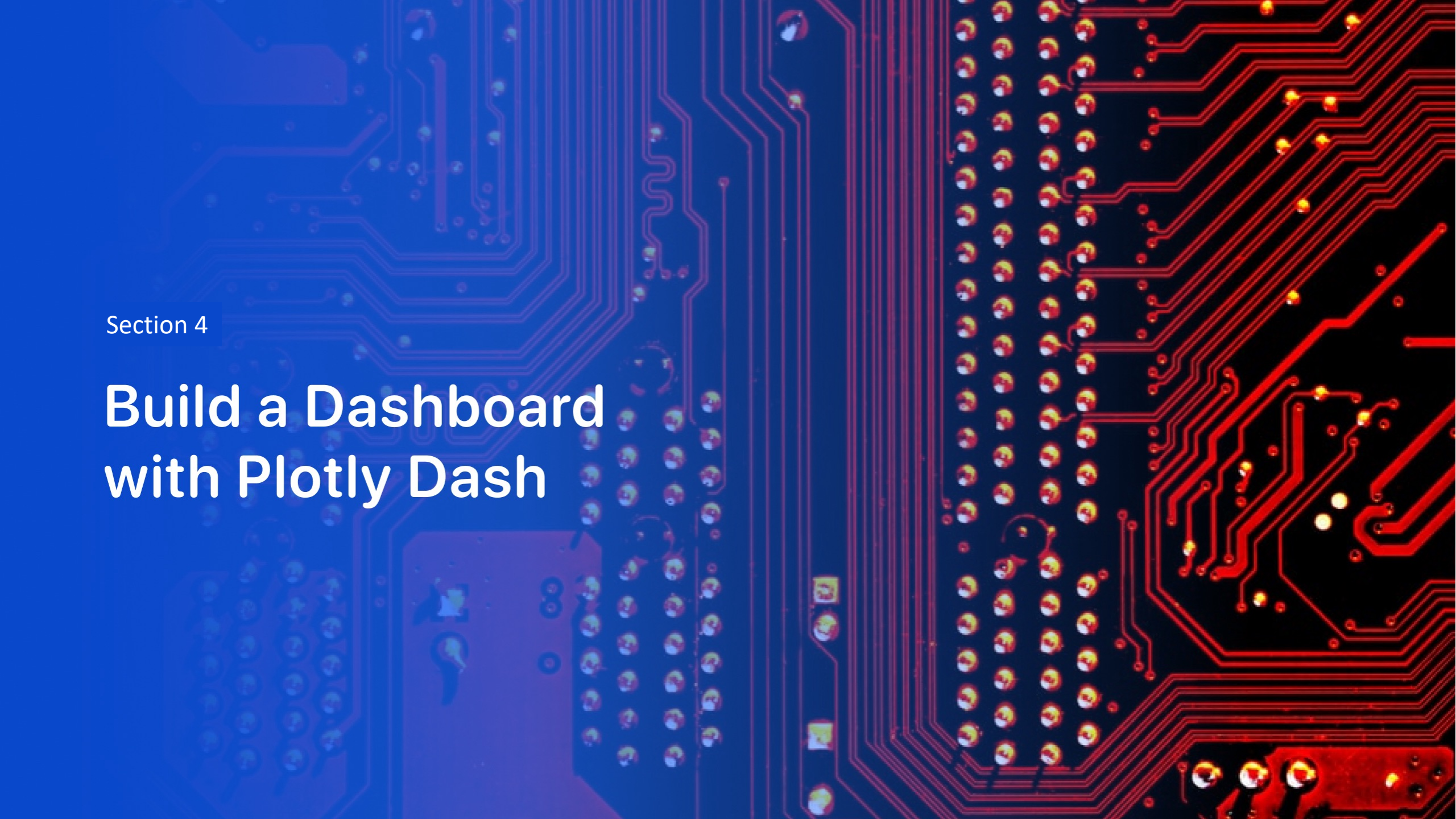
CCAFS LC 40

Folium Map: distance between CCAFS SLC-40 and proximities

The launch sites are relatively close to railways, highways and coastline compared to those from cities because of safety consideration

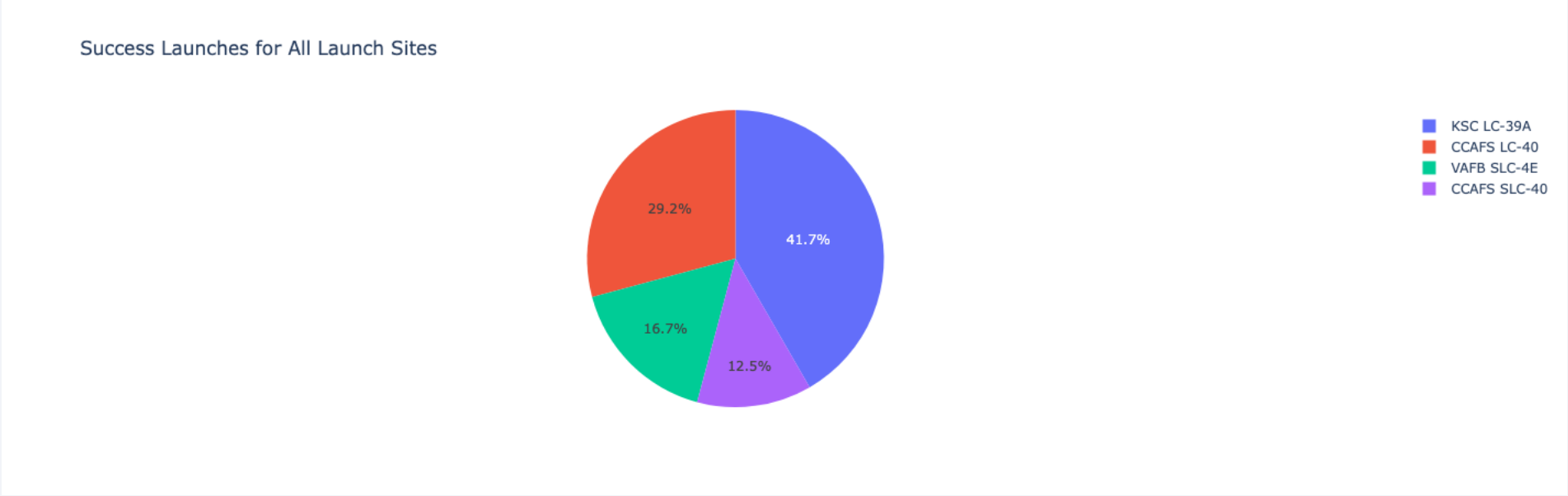
Moreover, the launch site is set at the optimal distance to railways, highways and coastline that they are not too close to the crowd, and at the same time, people concerned can approach launch sites at a reasonable time




Section 4
Build a Dashboard with Plotly Dash

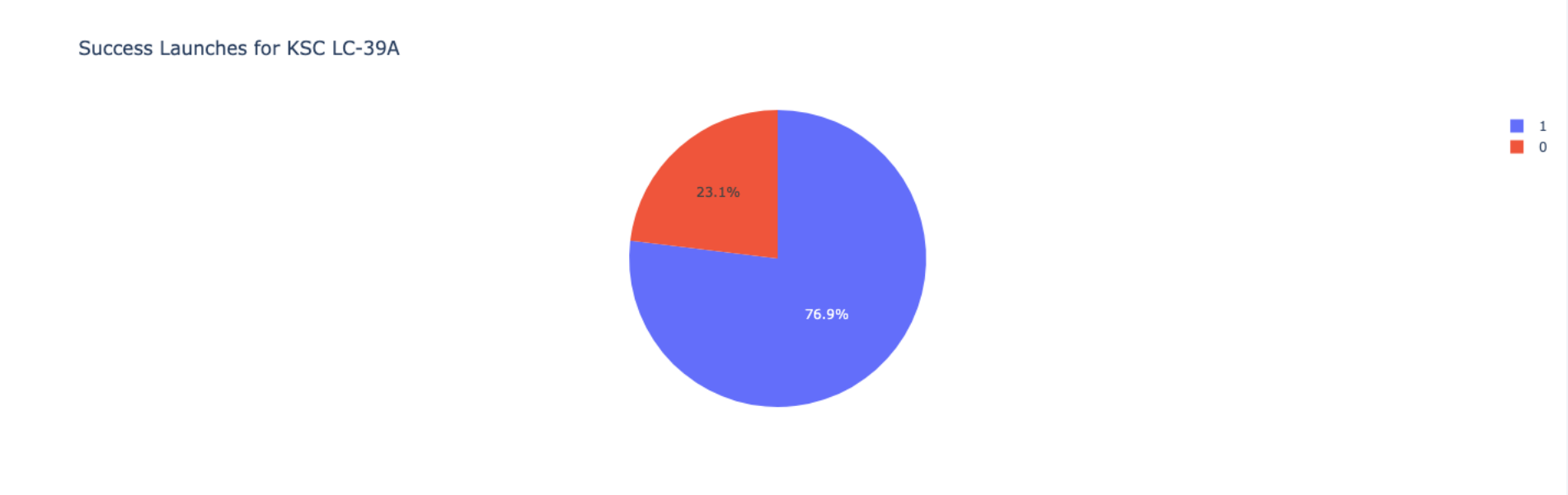
Dashboard: success launches for all launch sites

The success ratio of KSC LC-39A is the greatest among all launch sites, followed by CCAF SLC-40



Dashboard: success launch for KSC LC-40

For the launch site, KSC LC-40 with highest launch success ratio, the success launch ratio reaches 76.9%



Dashboard: payload vs. launch outcome

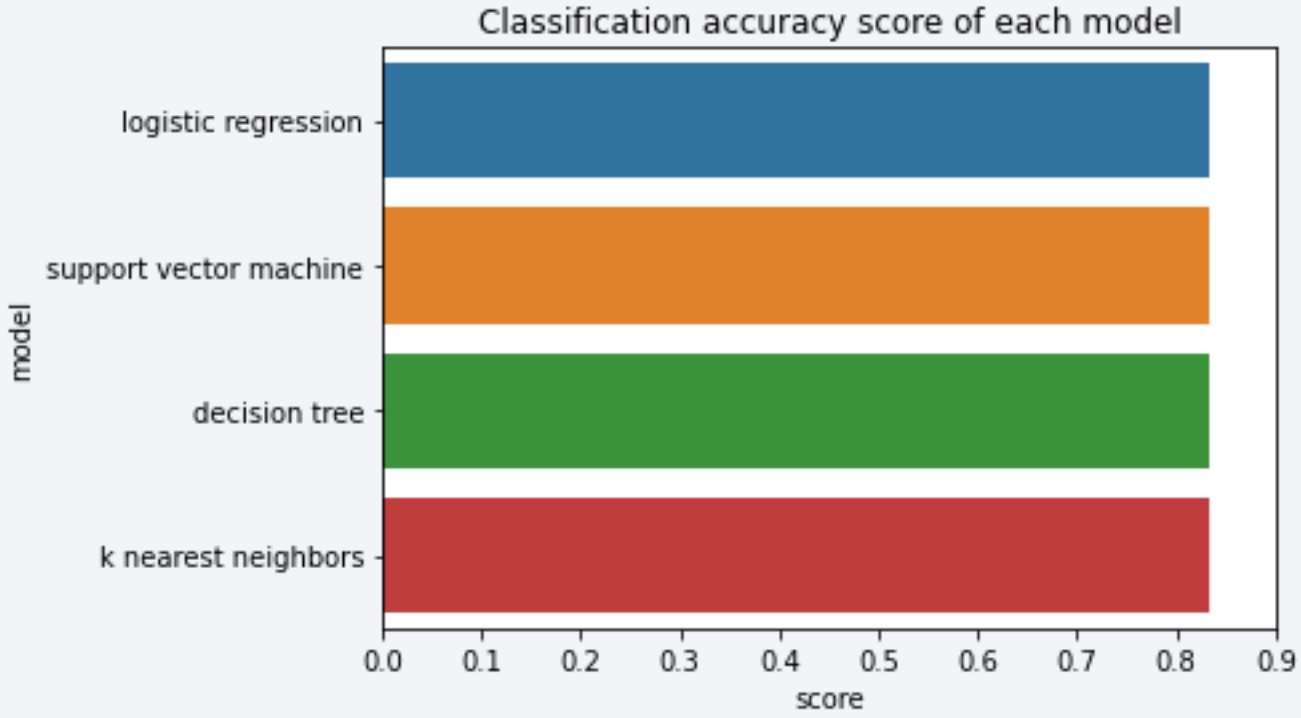
The success rate is the greatest when the payload mass is between 2,000 kg and 4,000 kg especially for FT booster version category

Section 5

Predictive Analysis (Classification)

Classification accuracy score

All models have the same classification accuracy of 0.833



Confusion Matrix

All models resulted in the same confusion matrix

Conclusions

- SpaceX Falcon 9 rocket launch data were requested via SpaceX API and scraped from Wikipedia, and the data was cleaned and wrangled.
- Exploratory data analysis was performed that the relationship between features was visualized and studied. There are four unique launch sites for Falcon 9 launch where KSC LC-39A have the greatest success launch. Moreover, ES-L1, SSO, HEO and GEO orbits have the greatest success launch. Furthermore, the success rate increases from 2013 to 2020.
- Important features including flight number, payload mass, orbit, launch site, flights, grid fins, reused, legs, landing pad, block, reused count and serial were selected, and classification models for Falcon 9 successful landing prediction were trained and evaluated with accuracy of 0.83.

Appendix

Notebook and Python script

Data Collection – SpaceX API

Data Collection – Scraping

Data Wrangling

EDA with Data Visualization

EDA with SQL

Interactive Map with Folium

Dashboard with Plotly Dash

Predictive Analysis

Thank you!