Winning Space Race with Data Science

Yixuan Wang
23 Dec 2021

Outline

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

Executive Summary

- Collect and process the data, obtain insights through exploratory data analysis (EDA) , train and test models to predict the Falcon 9 rockets success launching rate.
- In general increase in flight number, the success rate increases, high payload has low probability to launch successfully.
- The success launching rate differs for different boosters version, launch sites, orbit types.
- Success rate increases over the year.
- Machine learning can be developed to predict the rocket launching success rate, with an accuracy score higher than 80%.

Introduction

- SpaceX 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. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

- We will predict if the Falcon 9 first stage will land successfully.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Collecting data from SpaceX API

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

- Performing web scraping to collect Falcon 9 historical launch records from Wikipedia page, titled:

  List of Falcon 9 and Falcon Heavy launches

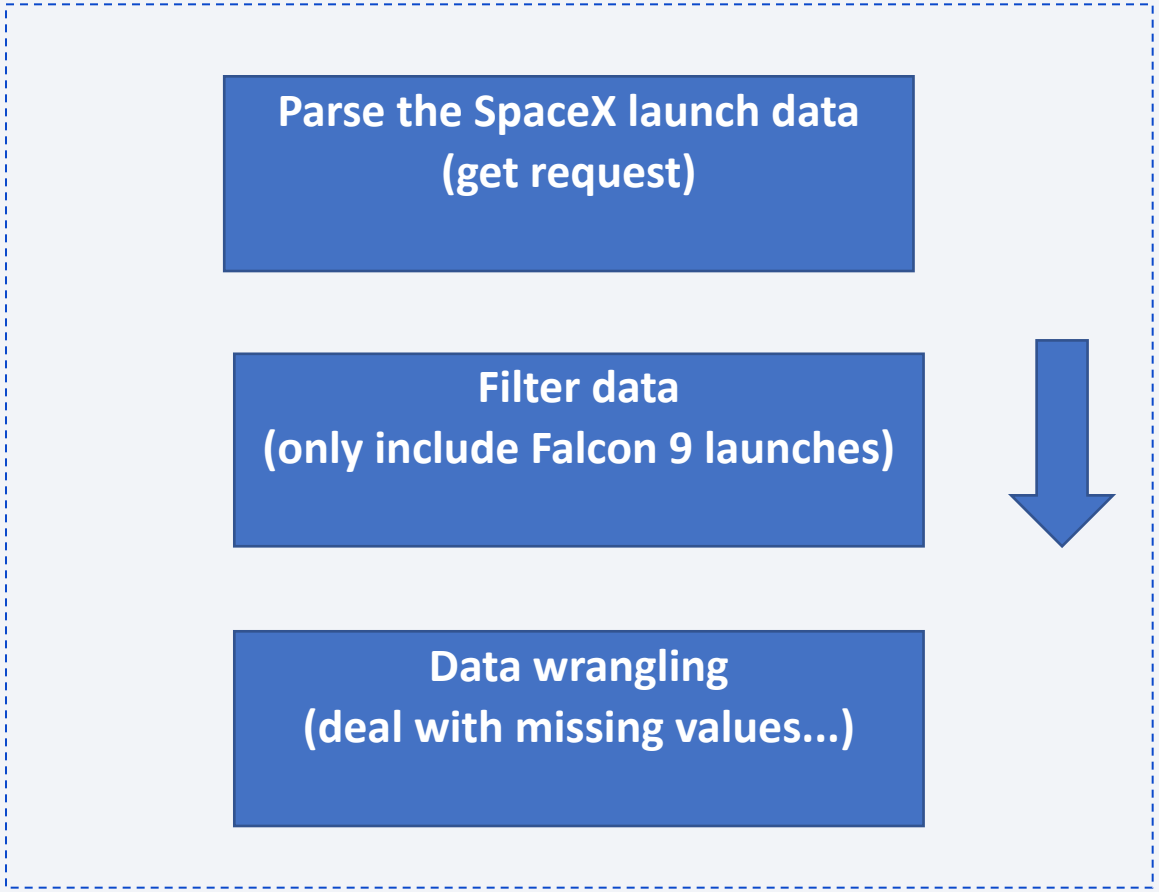
  https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Data Collection – SpaceX API

- Requesting rocket launch data from SpaceX API https://api.spacexdata.com/v4/launches/past

- GitHub URL for the detailed data transformation process:

  Data Collection with API

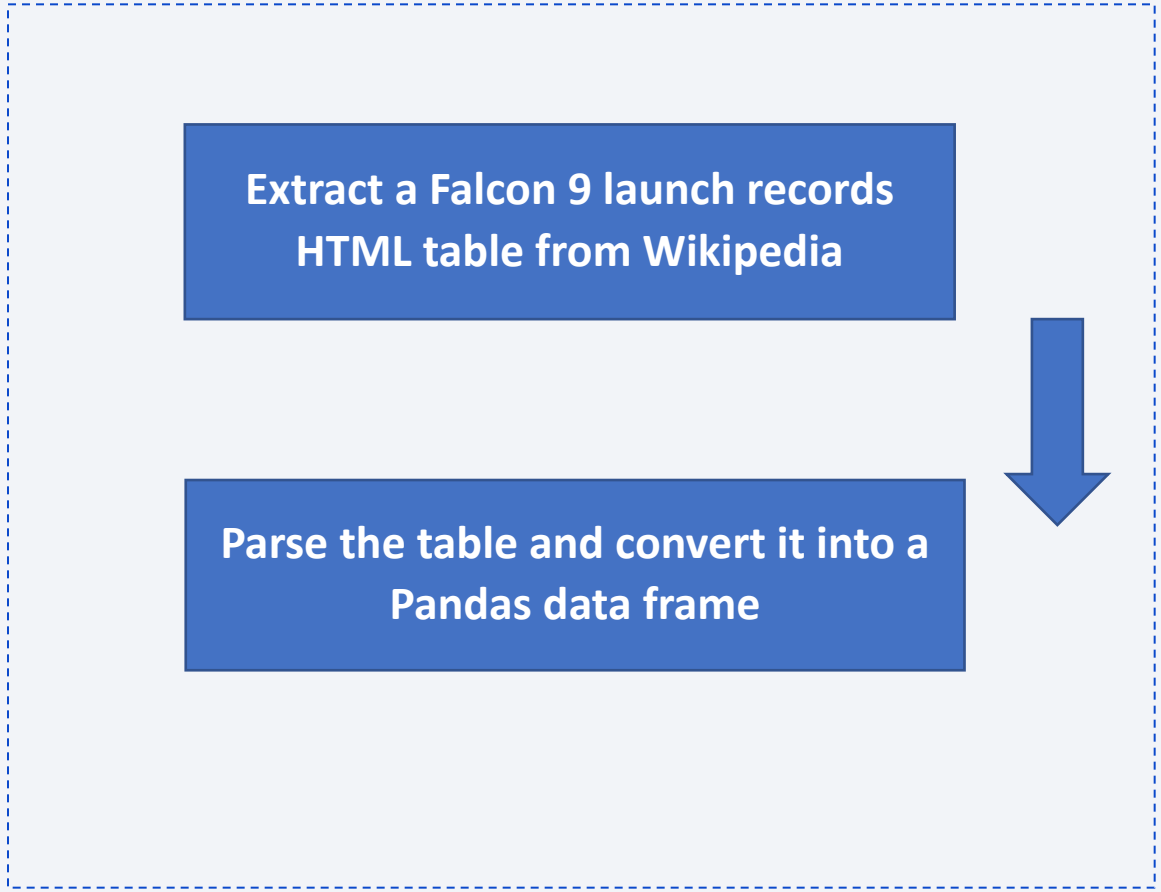


Data Collection - Scraping

- Web scraping Falcon 9 and Falcon Heavy Launches Records from Wikipedia

- GitHub URL for the detailed data transformation process:

  Data Collection withWeb Scraping



Data Wrangling

- Perform some Exploratory Data Analysis (EDA)
- Find some patterns in the data
- Determine what would be the label for training supervised models.

GitHub URL: Data Wrangling

Deal with missing values

Correct data format

Data standardization

Data normalization

Create dummy variables

EDA with Data Visualization

Following charts were plotted to find out how the success launch rate is influenced by the relevant factors:

1. Scatter plot of Flight Number vs. Payload
2. Scatter plot of Flight Number vs. Launch Site
3. Scatter plot of Payload vs. Launch Site
4. Bar chart for the success rate of each orbit type
5. Scatter point of Flight number vs. Orbit type
6. Scatter point of payload vs. Orbit type
7. Line chart of yearly average success rate

GitHub URL: EDA with Data Visualization

- Summarize what charts were plotted and why you used those charts
- Add the GitHub URL of your completed EDA with data visualization notebook, as an external reference and peer-review purpose

EDA with SQL

The SQL queries you performed for the exploratory data analysis:

- %sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXDATASET
- %sql SELECT * FROM SPACEXDATASET WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5
- %sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PADLOAD FROM SPACEXDATASET WHERE CUSTOMER='NASA (CRS)'
- %sql SELECT AVG(PAYLOAD_MASS__KG_) AS PAYLOAD_MASS__KG FROM SPACEXDATASET WHERE BOOSTER_VERSION= 'F9 v1.1'
- %sql SELECT MIN(DATE) FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (ground pad)'
- %sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXDATASET WHERE LANDING__OUTCOME='Success (drone ship)' AND (PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_<6000)

EDA with SQL

The SQL queries you performed for the exploratory data analysis:

- %sql SELECT LANDING__OUTCOME, COUNT(*) FROM SPACEXDATASET GROUP BY LANDING__OUTCOME
- %sql SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXDATASET WHERE PAYLOAD_MASS__KG_=(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET)
- %sql SELECT BOOSTER_VERSION, LAUNCH_SITE, DATE FROM SPACEXDATASET WHERE LANDING__OUTCOME ='Failure (drone ship)' AND(DATE LIKE'2015%')
- %sql SELECT LANDING__OUTCOME, COUNT(*) AS COUNT FROM SPACEXDATASET WHERE DATE>'2010-06-04' AND DATE<'2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY COUNT DESC

GitHub LInk: Exploratory Data Analysis with SQL

Build an Interactive Map with Folium

- Markers and circles folium object are used to plot all the launch site locations.
- Marker clusters can be a good way to simplify a map containing many markers having the same coordinate.
- With color-labeled markers in marker clusters, we identify which launch sites have relatively high success rates.
- Add a MousePosition on the map to get coordinate for a mouse over a point on the map. As such, we can find the coordinates of any points of interests.
- Draw a line and indicate the distance between a launch site to its closest highway, coast and so on.

GitHub URL: Launch Sites Locations Analysis with Folium

Build a Dashboard with Plotly Dash

- To understand the relationship between the success launch rate and the different launch sites, we create interactive pie chart to visualize overall success launches from different sites and there success rate respectively.

❖ Pie chart of the success launches overview by sites

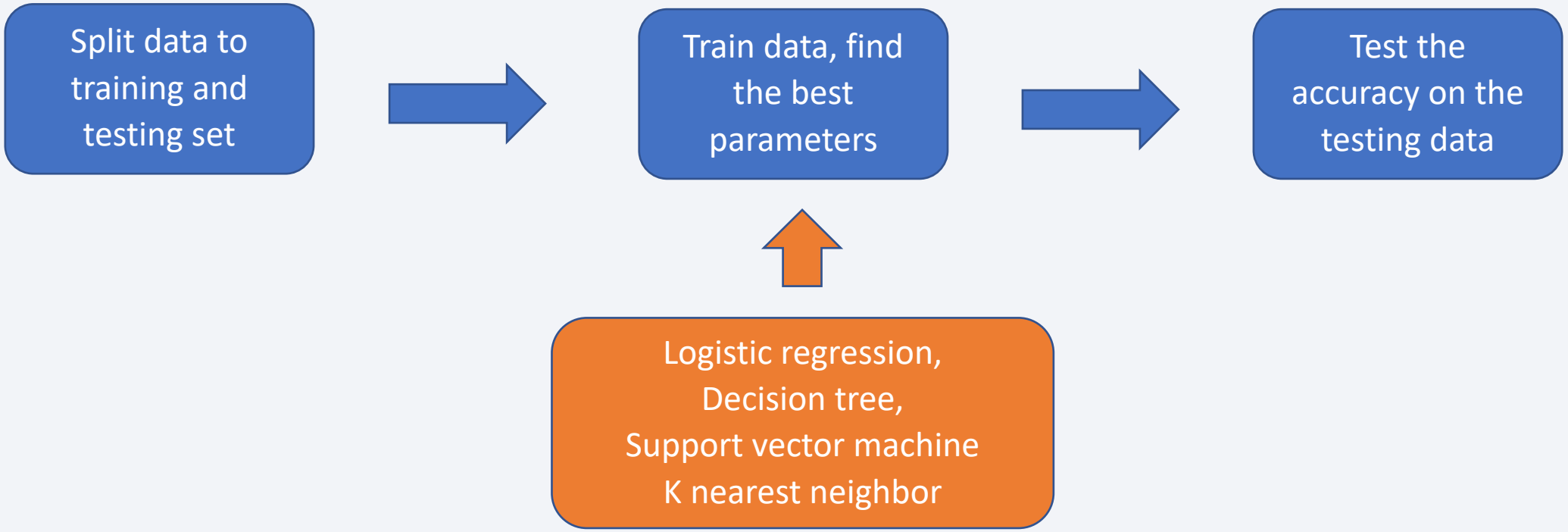
❖ Pie chart of the launch site with the highest launch success rate

❖ …

- Create interactive scatter plot with slide bar to explore the relationship between different payload with the success launch rate.

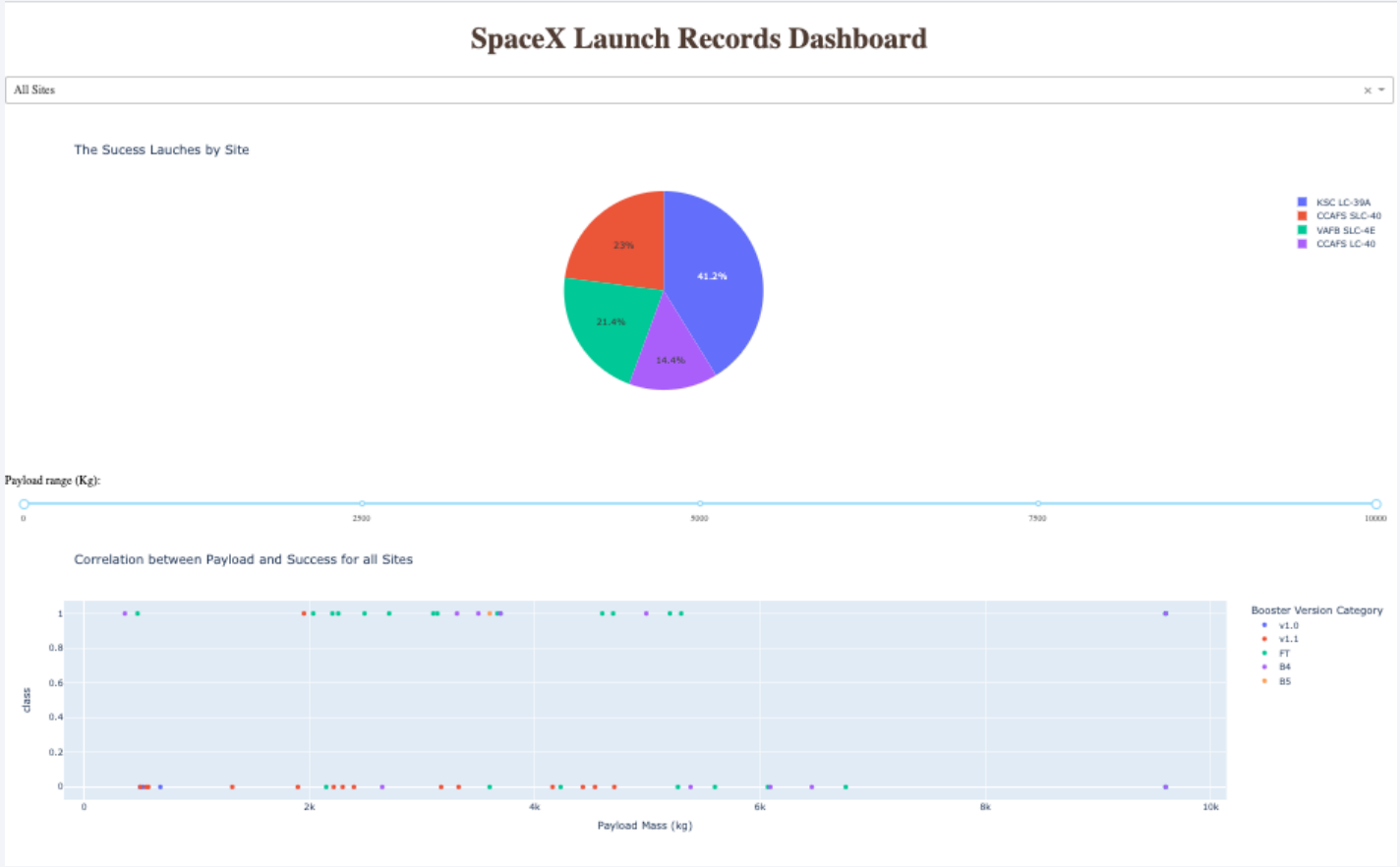
GitHub URL: Build a Dashboard Application with Plotly Dash

Predictive Analysis (Classification)



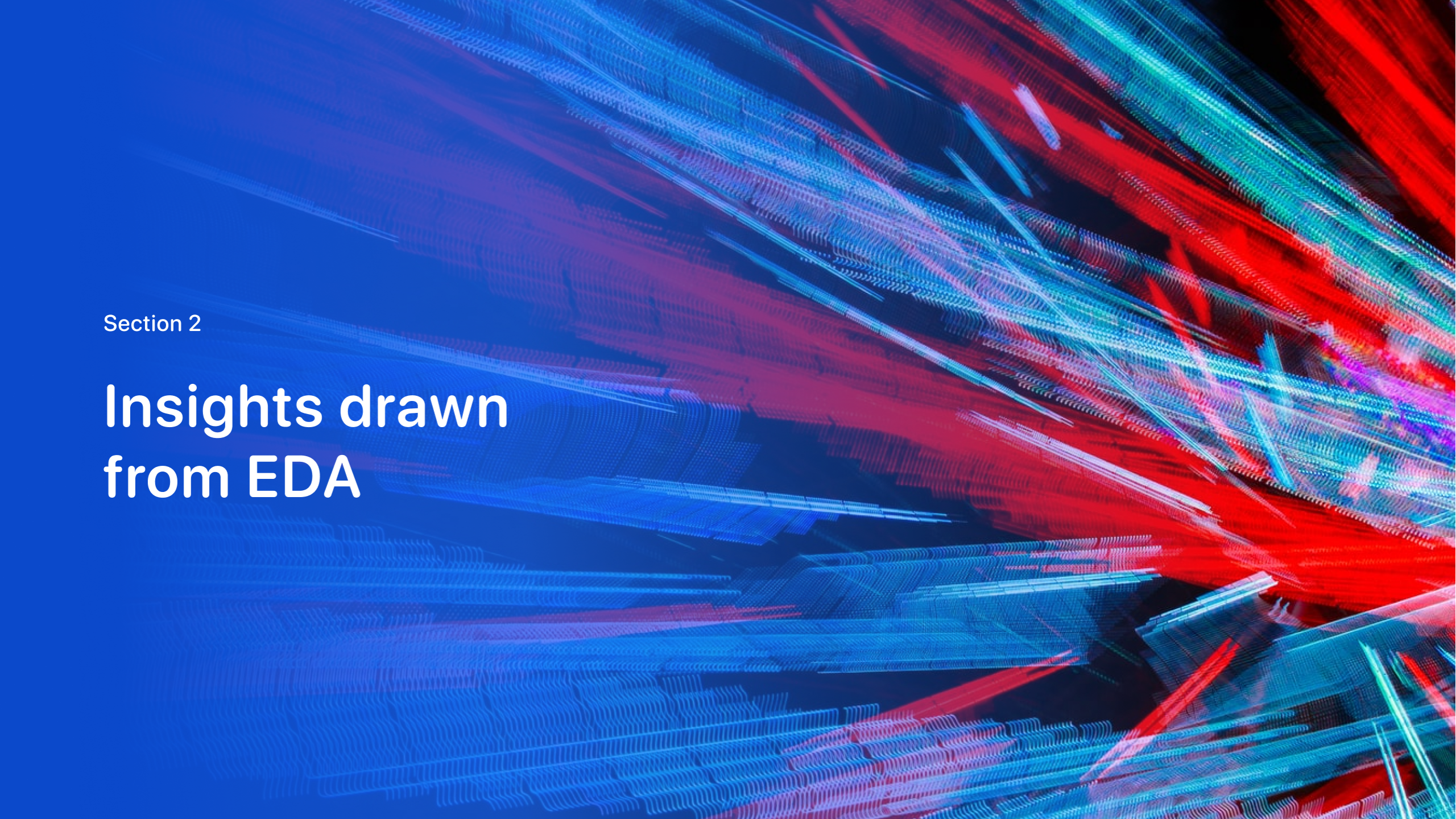
GitHub URL: Machine Learning Prediction

Results



- In general increase in flight number, the success rate increases, high payload has low probability to launch successfully. The success launching rate differs for different boosters version, launch sites, orbit types. Success rate increases over the year.

- Machine learning can be developed to predict the rocket launching success rate, with an accuracy score higher than 80%.

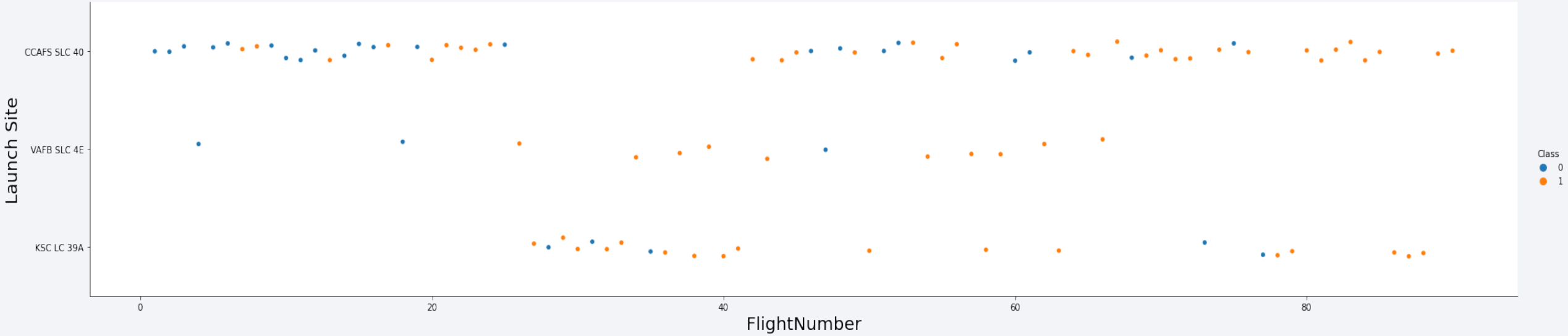


Section 2

Insights drawn from EDA

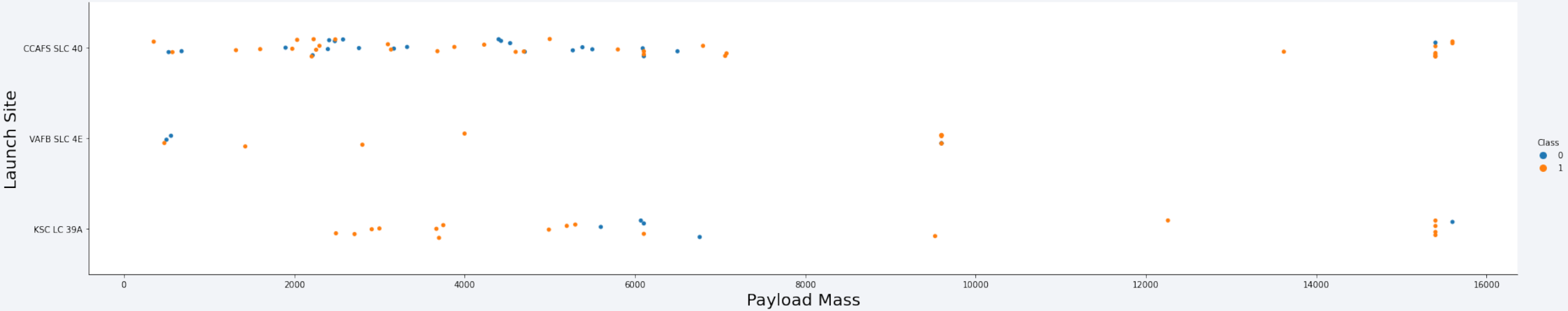
Flight Number vs. Launch Site

- FlightNumber: indicating the continuous launch attempts.
- Launch sites : CCAFS LC-40, KSC LC-39A and VAFB SLC 4E.



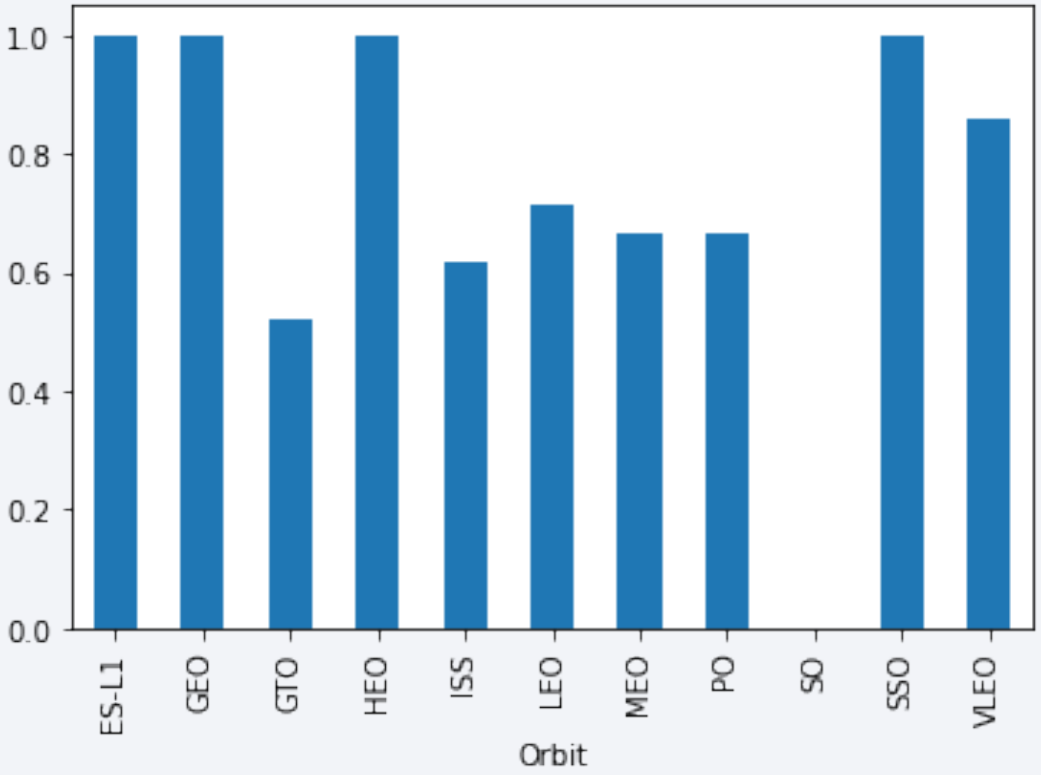
Hight Fight number can increase the success rate, and "VAFB SLC 4E" appears to have relatively high success rate and the most experiment carried out in CCAFS SLC 40.

Payload vs. Launch Site



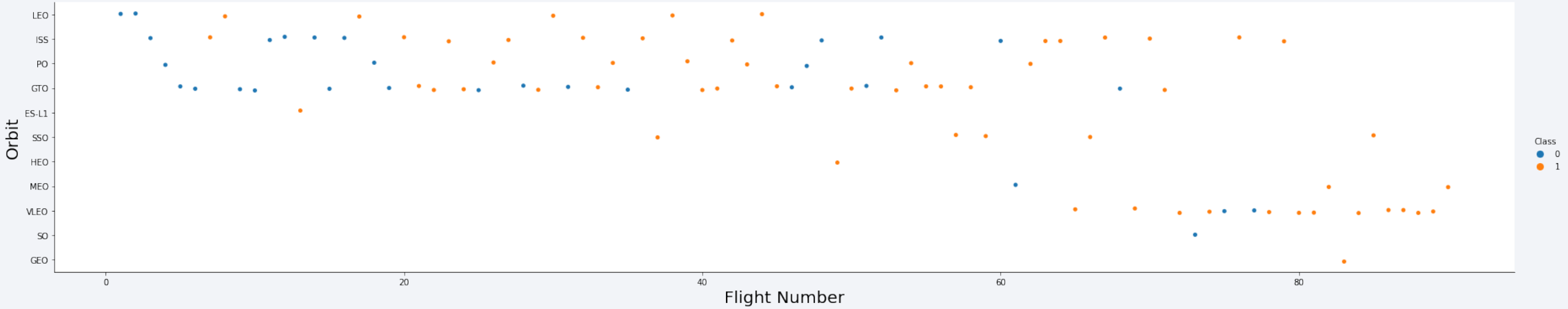
For the VAFB-SLC launchsite there are no rockets launched for heavypayload mass (greater than 10000).

Success Rate vs. Orbit Type



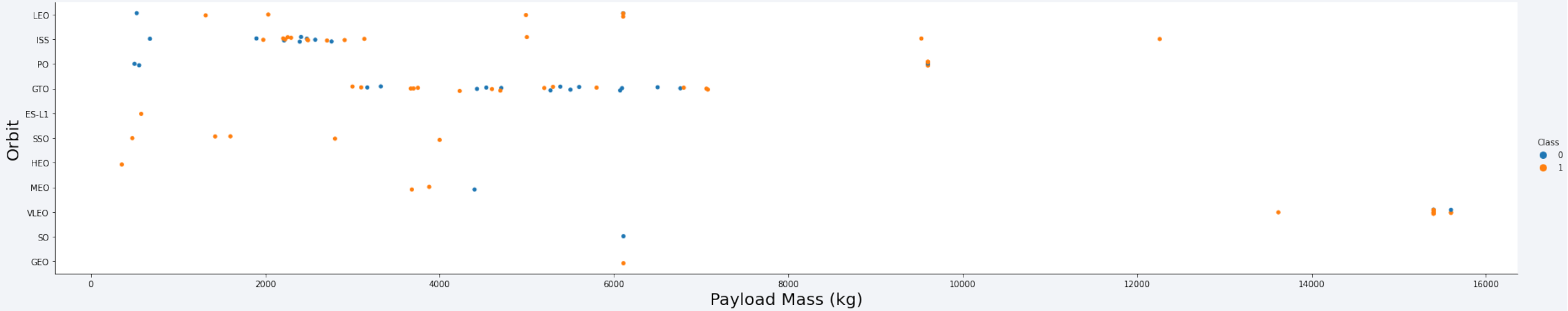
ES-L1, GEO, HEO and SSO orbit had the highest success rate, 100%

Flight Number vs. Orbit Type



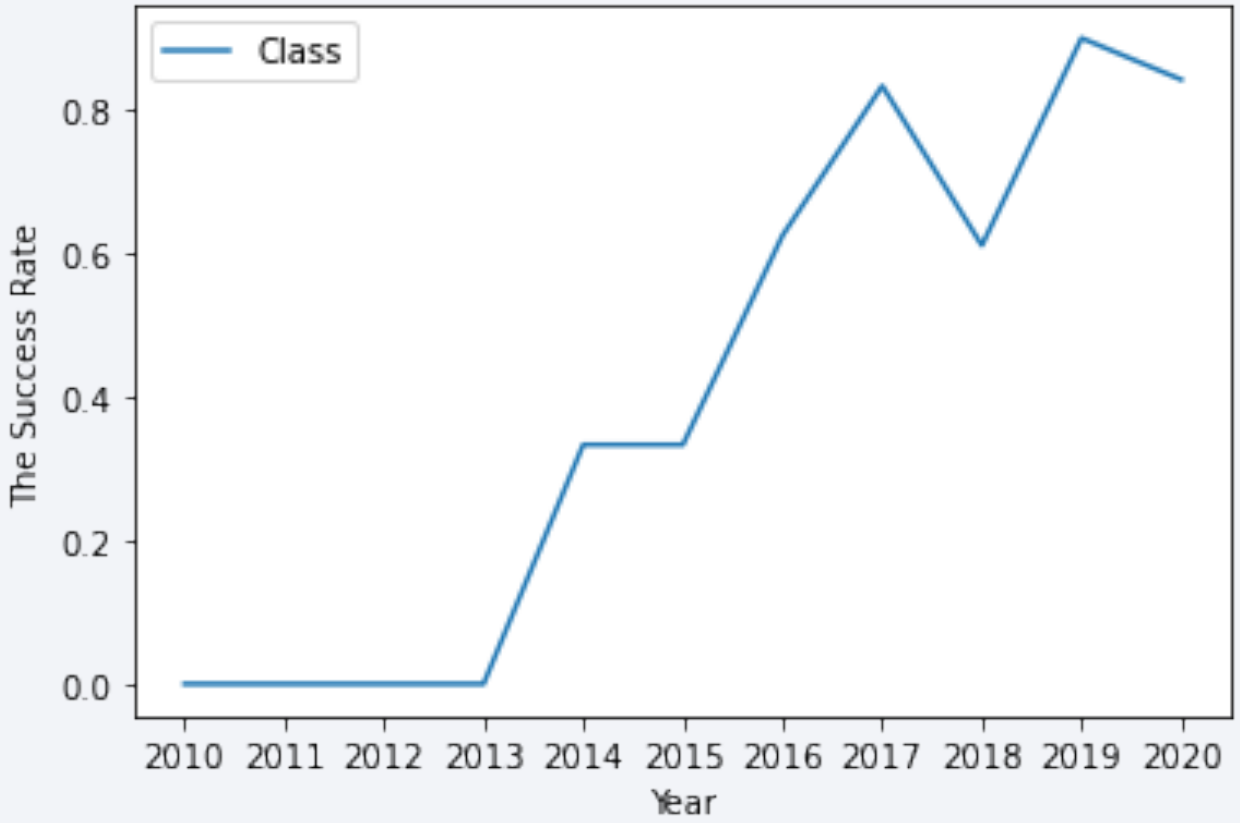
- LEO orbit the Success appears related to the number of flights;
- On the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there.

Launch Success Yearly Trend



The sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

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

- There have been four launch sites for SpaceX Falcon 9

Launch Site Names Begin with 'CCA'

- Launch sites begin with `CCA` which can be CCAFS LC-40 or CCAFS SLC-40

  (First 5 records sorted by date)

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

total_padload
45596

- Calculate the total payload carried by boosters from NASA (CRS)
- Unit: Kg

Average Payload Mass by F9 v1.1

payload_mass__kg
2928

The average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

	1
	2015-12-22

The dates of the first successful landing outcome on ground pad.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

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

Total Number of Successful and Failure Mission Outcomes

landing__outcome	2
Controlled (ocean)	5
Failure	3
Failure (drone ship)	5
Failure (parachute)	2
No attempt	22
Precluded (drone ship)	1
Success	38
Success (drone ship)	14
Success (ground pad)	9
Uncontrolled (ocean)	2

The total number of successful and failure mission outcomes from 2010 to 2020.

Boosters Carried Maximum Payload

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

The names of the booster which have carried the maximum payload mass, 15600 kg.

2015 Launch Records

booster_version	launch_site	DATE
F9 v1.1 B1012	CCAFS LC-40	2015-01-10
F9 v1.1 B1015	CCAFS LC-40	2015-04-14

List of the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

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

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

Rank of 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

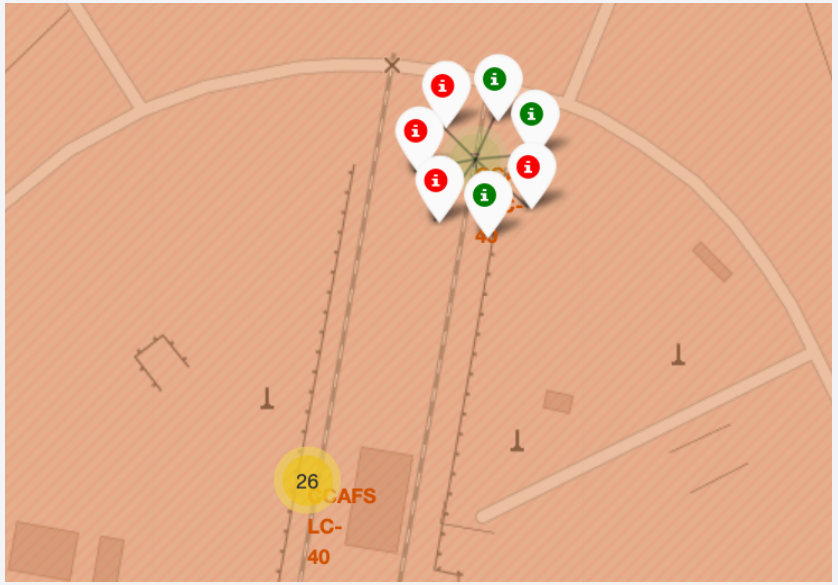
Section 4

Launch Sites Proximities Analysis

All Launch Sites' Location Markers on a Global Map

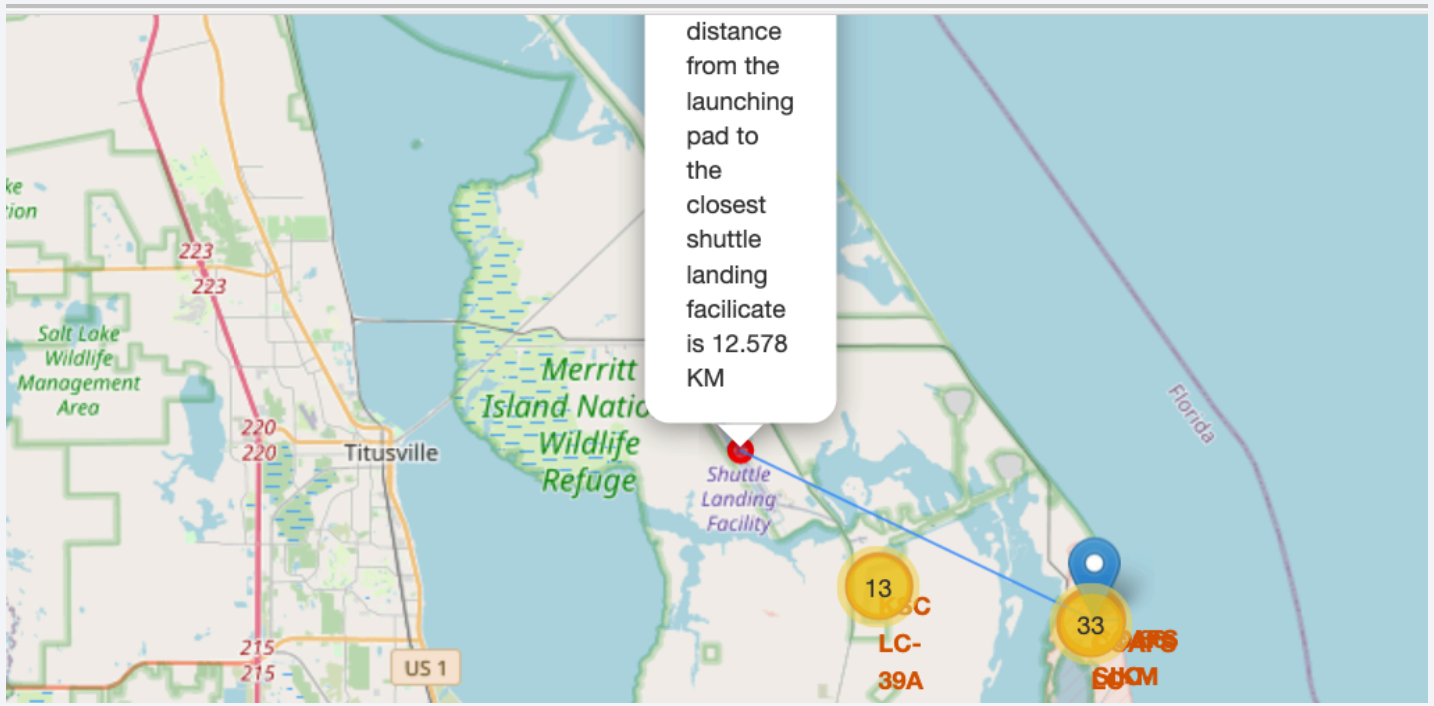
- All launch sites are in proximity to the Equator line.
- All launch sites are very close proximity to the coast.

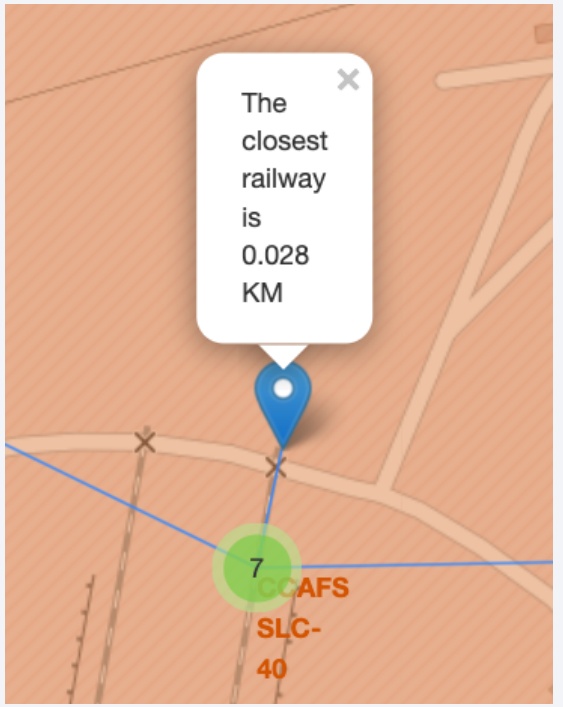
Launch Sites Map with Success/Failed Markers



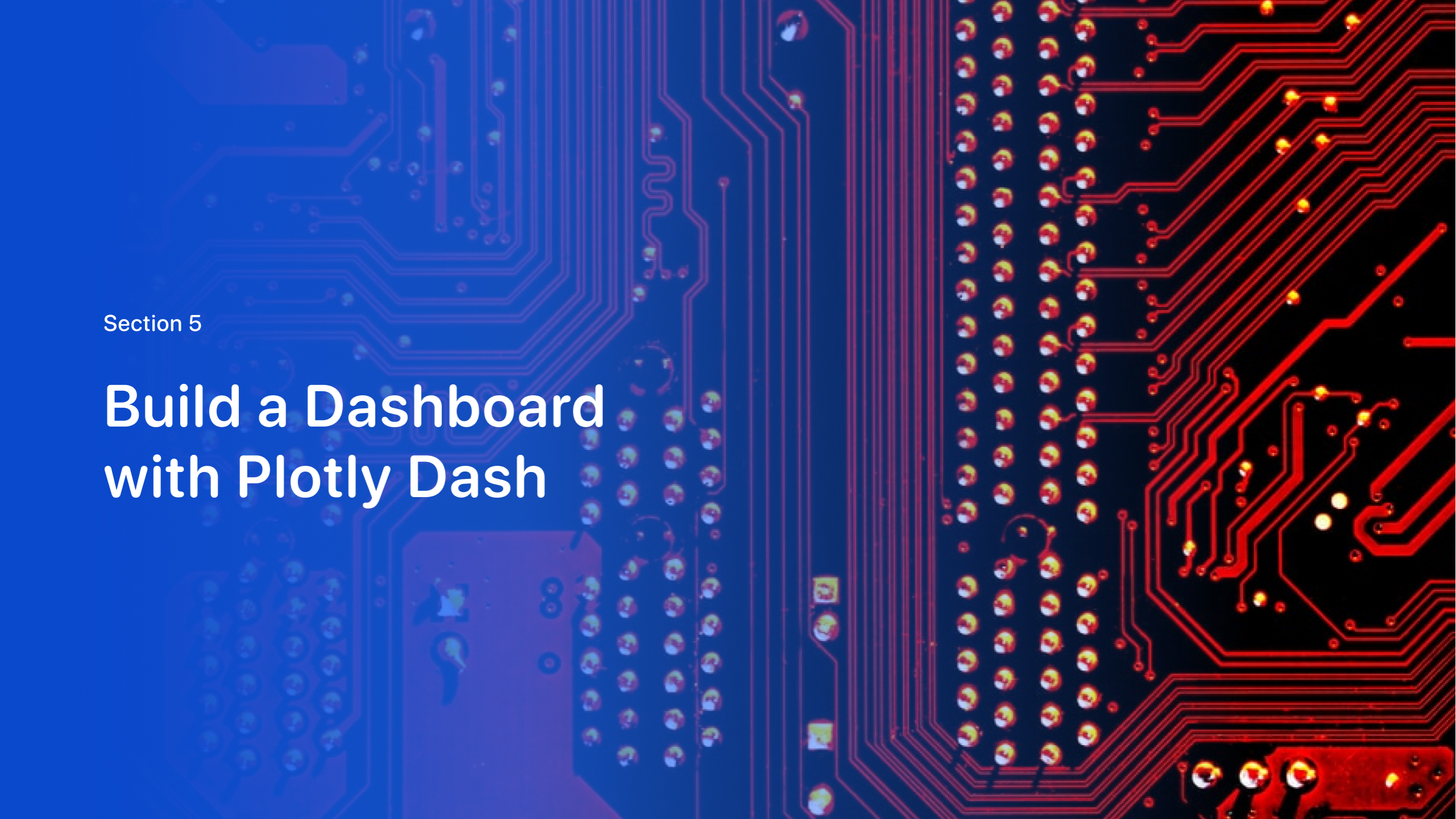
From the color-labeled markers in marker clusters, we are able to identify the success rates of each launch site.

Exploring the Environment around Launch Site





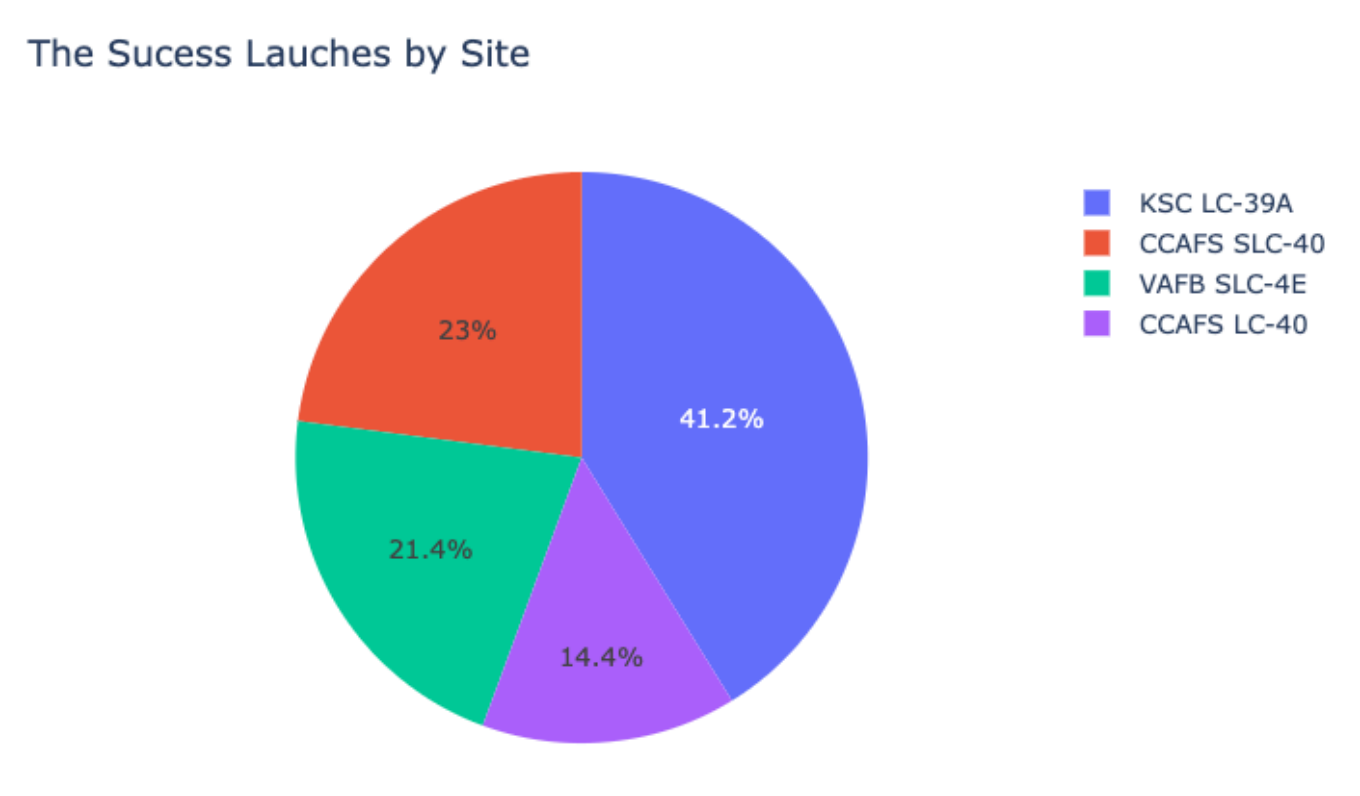
The distance of the CCAFS SLC-40 launch site to its closest shuttle landing facilitate is 12.6km and to the closest railway is only 28 m.



Section 5

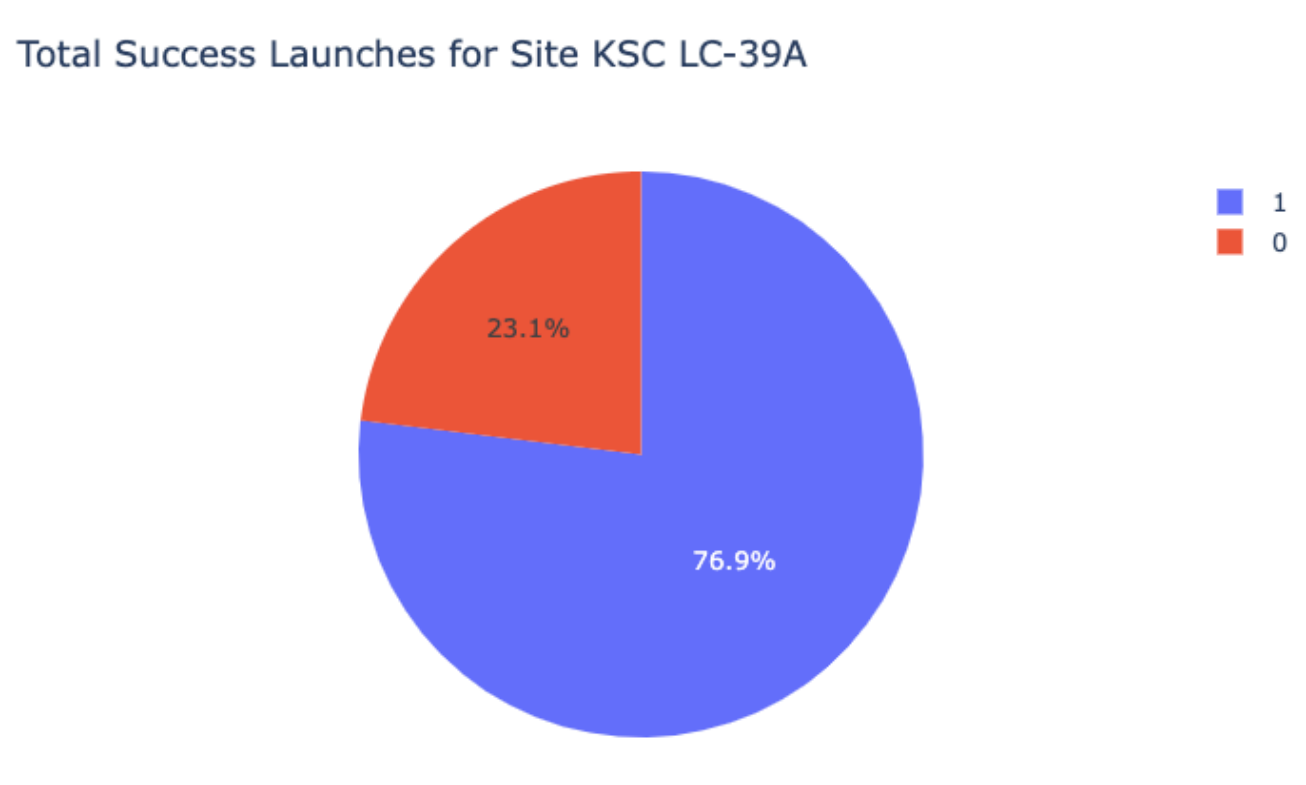
Build a Dashboard with Plotly Dash

The Success Launches Overview by Sites



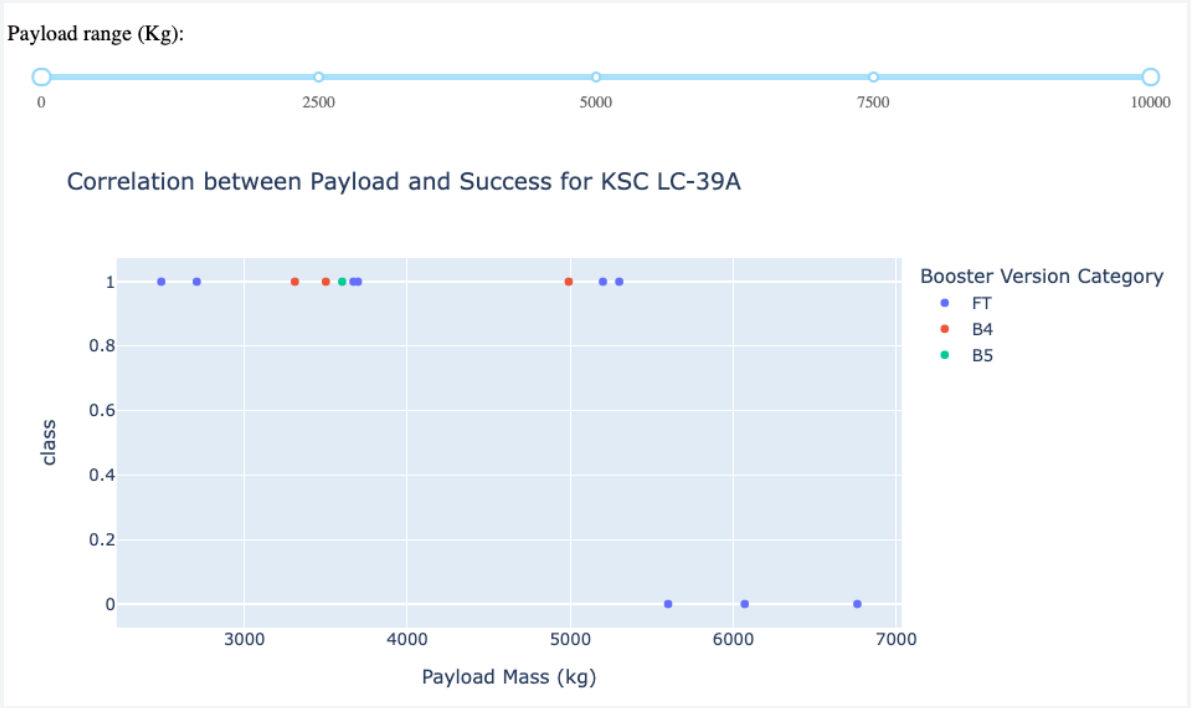
The KSC LC-39A launch site has the most success launches, 41.2%.

Launch Site with the Highest Launch Success Rate



The KSC LC-39A launch site has the highest success launch rate, 76.9 %.

Different Payloads vs. Launch Outcome Scatter Plot



- When payload is higher than 5500 kg, the success rate drops to 0.
- On the other hand, launches with payload less than 5500 kg succeeded in the past.
- The high payload launches were only experimented with the FT Booster.

Section 6

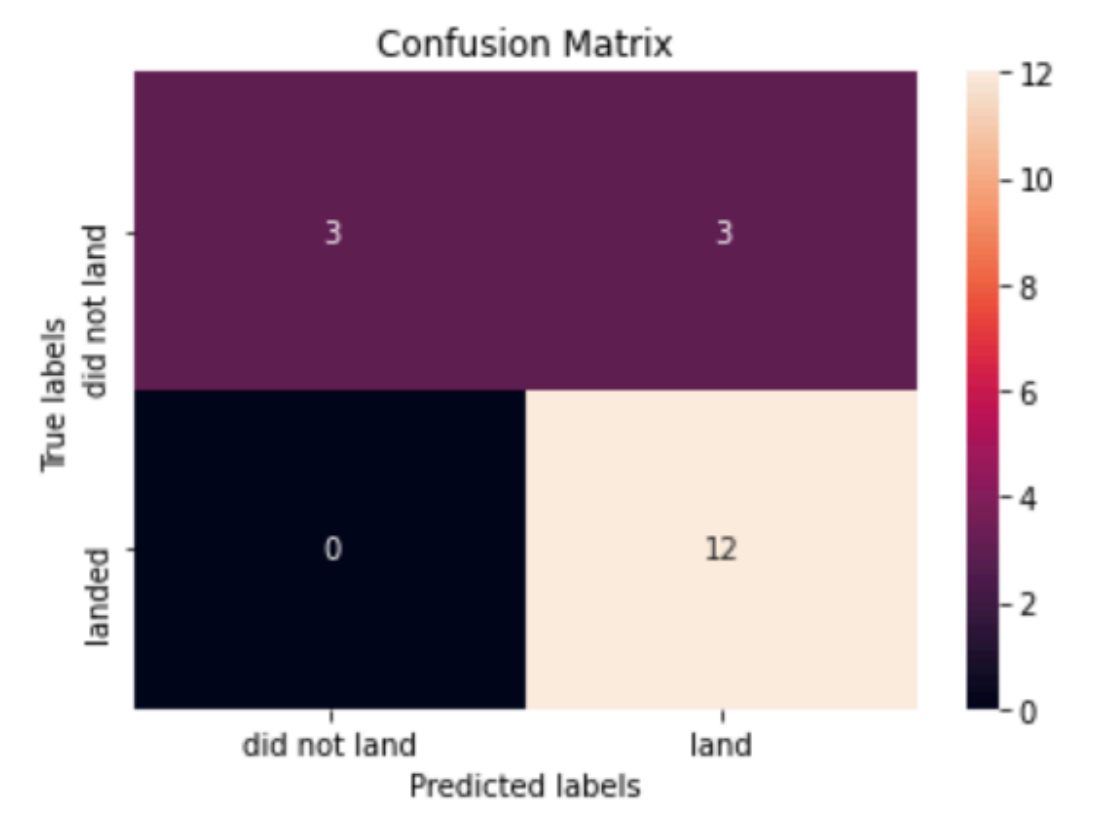
Predictive Analysis (Classification)

Classification Accuracy

Model	Accuracy for training data	Accuracy for testing data
logistic regression	0.846	0.833
support vector machine	0.848	0.833
decision tree classifier	0.889	0.833
k nearest neighbors	0.848	0.833

- For the training data, the decision tree model has a better score
- For testing set, the classification accuracy is the same, 0.833.

Confusion Matrix



- When testing the data, all models have the same accuracy score and have the same confusion matrix.
- The confusion matrix implies that the major problem for all the models is the false positives.

Conclusions

- Booster version, payload mass, launch site .etc influence the launching results.
- Machine learning can be used to predict the success launching rate for the SpaceX's Falcon 9 rocket, with a relatively high accuracy rate, 0.833.
- False positives occur in all models: when the model predicts a successful land while actually results in a failure.

Appendix

Machine learning data sample without processing:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
0	1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
1	2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
2	3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
3	4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
4	5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0

GitHub URL: Machine Learning Prediction

Thank you!