IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Aparajita Jha
12/07/2022

Outline

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

Executive Summary

- Methodologies:
- Collected data from public SpaceX API and SpaceX Wikipedia page.
- Explored data using SQL, visualization, folium maps, and dashboards.
- Standardized data and used GridSearchCV to find best parameters for machine learning models.
- Visualized accuracy score of all models.

- Results:
- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors that provided similar results with accuracy rate of about 83.33%.
- All models predicted successful landings.

Introduction

- Project background and context
- Space X, headed by Elon Musk has been pivotal due to its ability for producing effective rockets.
- Aim of the project is to establish an enterprise, Space Y, to compete with Space X.

- Problems you want to find answers
- To train Machine Learning Model to predict successful landing.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
- Perform data wrangling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

Space X API Data Columns:

- FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

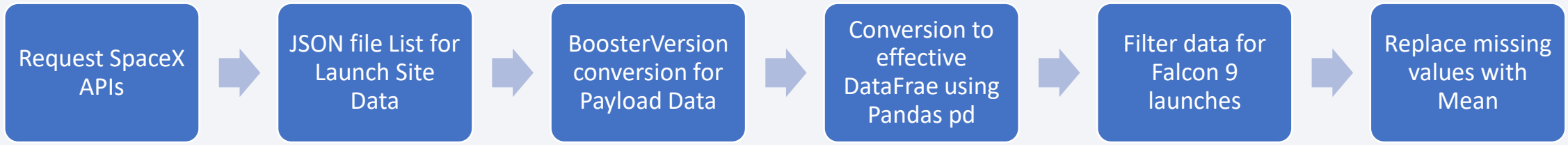
Wikipedia Webscrape Data Columns:

- Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

GitHub URL of the completed SpaceX API

https://github.com/AJ-DarKnight/IBM_DataScience/blob/main/1_Data%20Collection.ipynb



Data Collection - Scraping

GitHub URL of the completed Data Collection using Scraping and EDA

https://github.com/AJ-DarKnight/IBM_DataScience/blob/main/2_EDA.ipynb

Data Wrangling

Exploratory data analysis was used to uncover certain trends in the data and choose the label for supervised model training.

There were several instances of the booster failing to effectively land in the data set. Sometimes a landing attempt fails due to an accident; for instance,

- True Ocean indicated that a mission's outcome was a successful landing in a certain ocean location whereas False Ocean indicated that a mission's end was an unsuccessful landing in a particular ocean zone.
- True RTLS indicated that the mission was accomplished when a ground pad was reached. False RTLS resulted in a failed landing on a ground pad for the mission.
- True ASDS meant that the drone ship was successfully landed at the end of the mission. False ASDS indicated that the drone ship landing attempt had failed.

Here, our major goal was to translate those results into training labels, where 1 denoted a successful booster landing and 0 indicated a failure.

EDA with Data Visualization

Analysis of exploratory data was done on the factors Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year.

- The following plots were used:

Flight Number vs Payload Mass, Flight Number vs Launch Site, Payload Mass vs Launch Site, Orbit vs Success Rate, Flight Number vs Orbit, Payload vs Orbit, and Success Yearly Trend.

In order to determine whether a link between the variables existed so that they could be employed in training the machine learning model, scatter plots, line charts, and bar graphs were used to compare relationships between the variables.

GitHub URL of EDA with Data Visualization:

https://github.com/AJ-DarKnight/IBM_DataScience/blob/main/4_EDAwithVisualization.ipynb

EDA with SQL

- Loaded dataset to Database using sqlite by connecting to my_data1.db due to IBM internal server error 500.
- Queried using %sql Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

GitHub URL of EDA with SQL:

https://github.com/AJ-DarKnight/IBM_DataScience/blob/main/3_EDASQL.ipynb

Build an Interactive Map with Folium

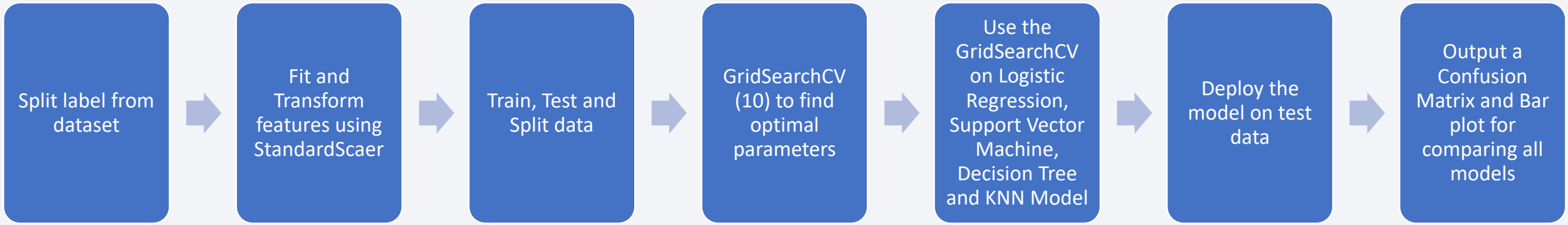
- Launch sites, successful and unsuccessful landings, and examples of important destinations that are close by are marked on folium maps, including railways, highways, coasts, and cities.
- This enables us to comprehend potential reasons for the positioning of launch locations that visualises successful landings in relation to their location as well.
- GitHub URL of EDA with Data Visualization:

https://github.com/AJ-DarKnight/IBM_DataScience/blob/main/5_Interactive%20Visual%20Analytics%20with%20Folium%20LAB.ipynb

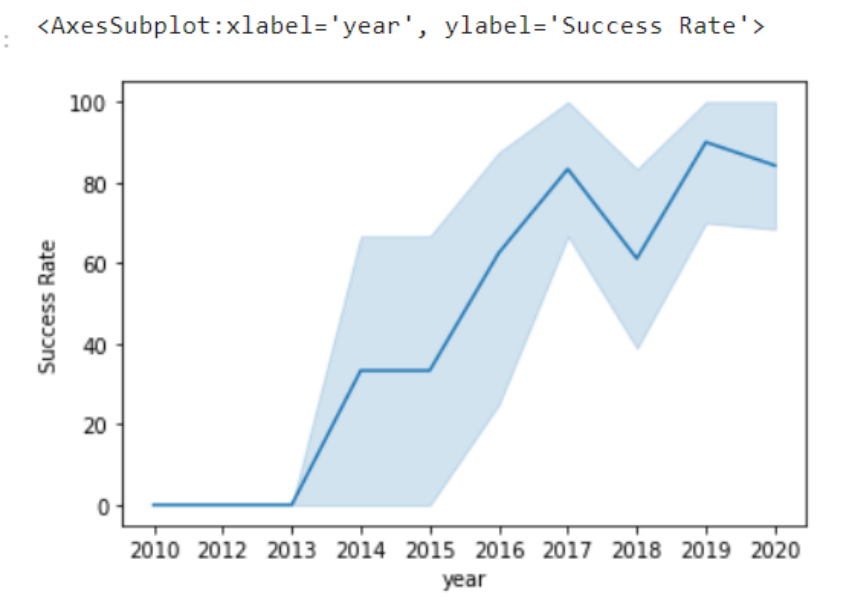
Predictive Analysis (Classification)

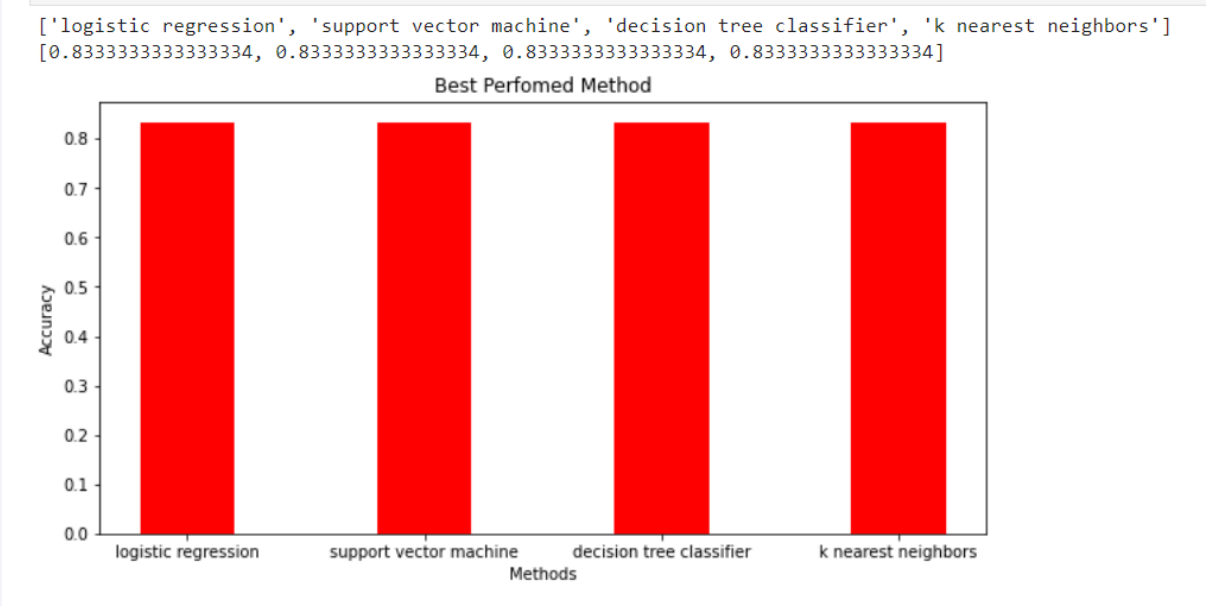
- GitHub URL of Classification:

https://github.com/AJ-DarKnight/IBM_DataScience/blob/main/6_Machine%20Learning%20Prediction.ipynb

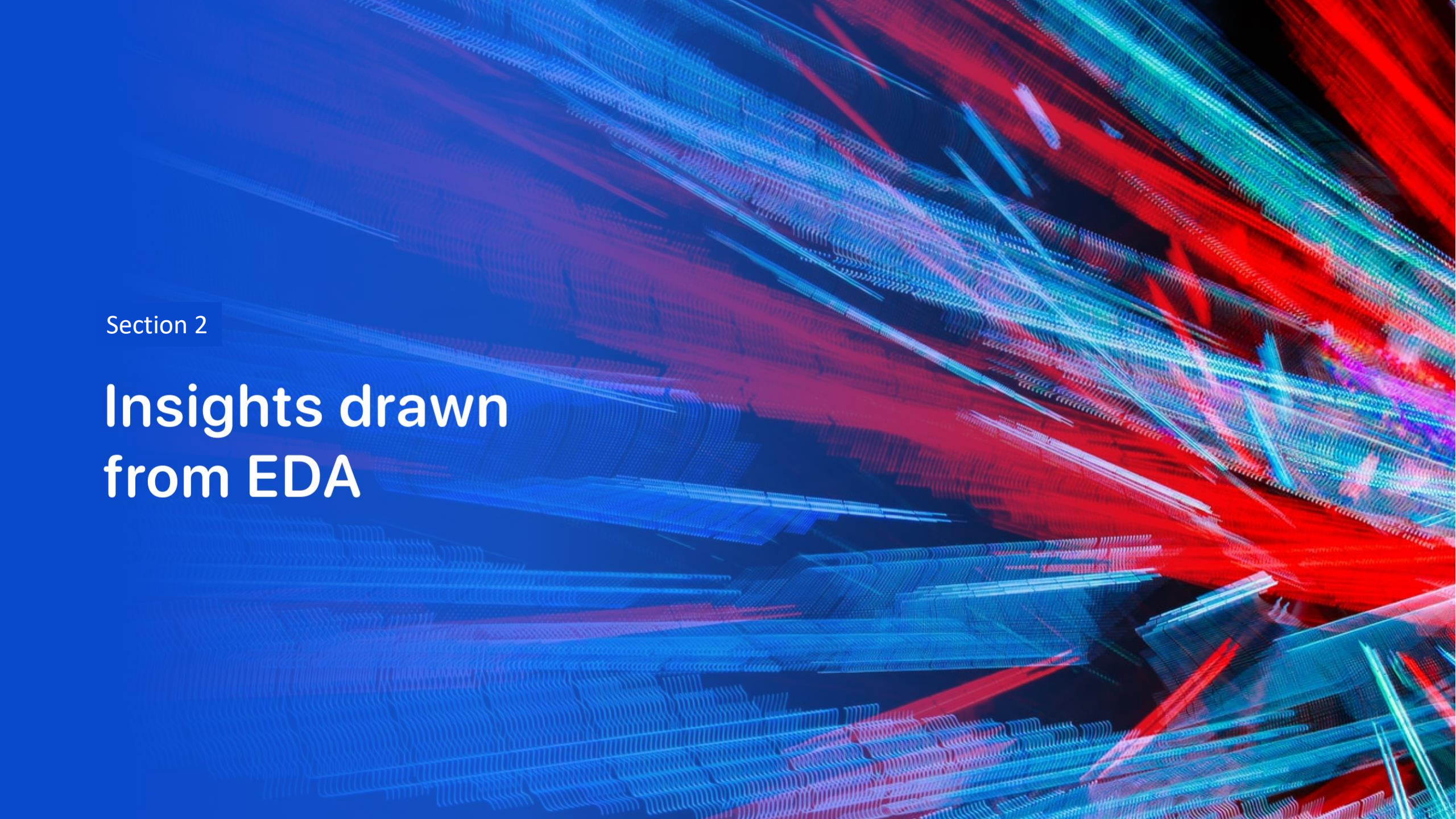


Results

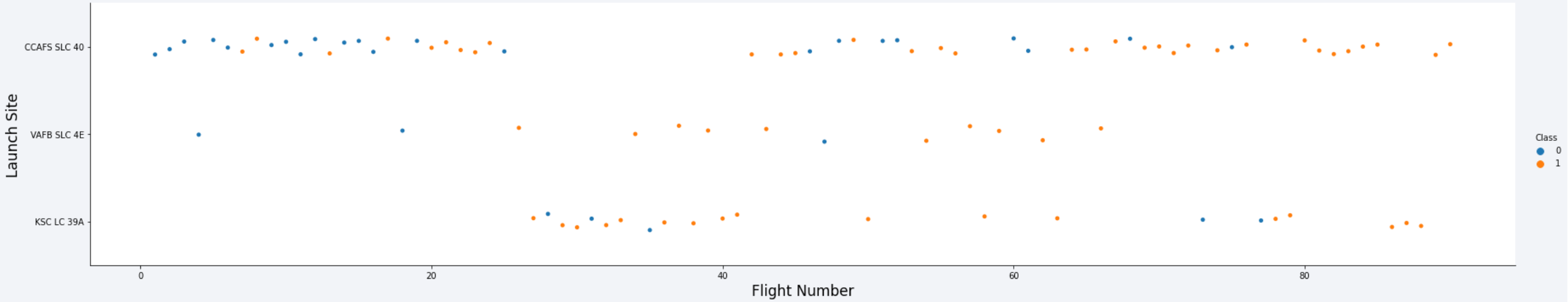




Based on EDA, one can observe that the sucess rate of landing since 2013 kept increasing till 2020 thus a method that performs best using all the hyperparameters was needed where all models resulted in ~83 accuracy for the SpaceY model to be deployed..


Section 2
Insights drawn from EDA

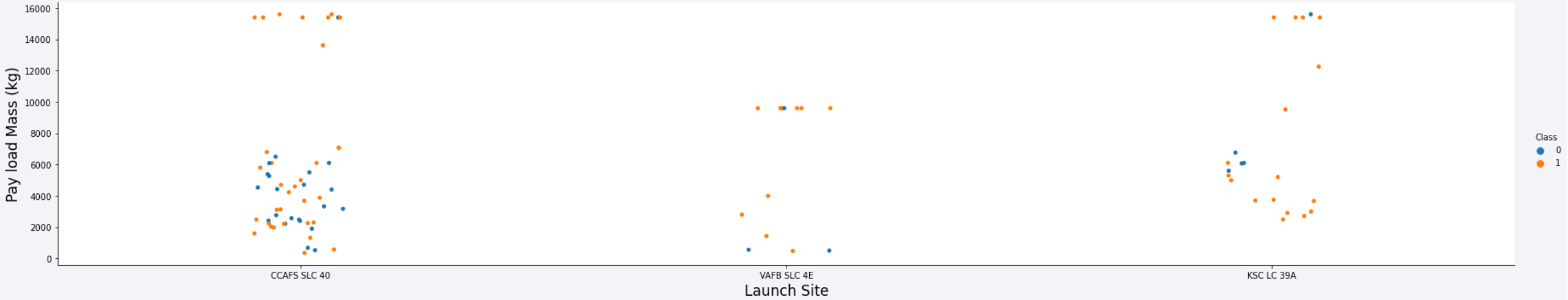
Flight Number vs. Launch Site



Blue states unsuccessful launches while yellow states successful launch.

The graphic hints to a rising success rate over time, as seen by the flight number. Most likely, there was a huge development around flight 20 that greatly improved the success rate. Given its volume, CCAFS looks to be the primary launch point.

Payload vs. Launch Site



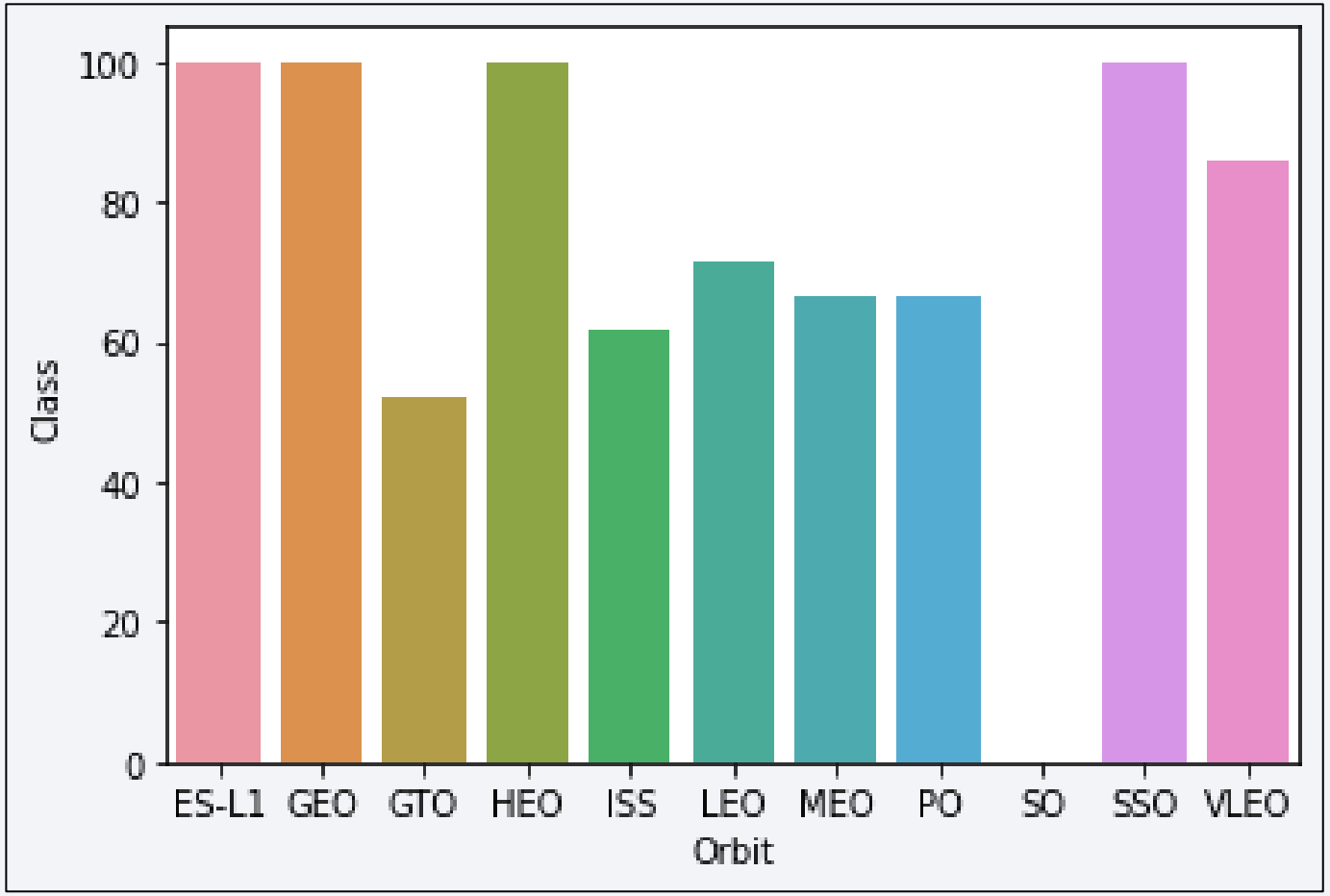
Blue states unsuccessful launches while yellow states successful launch.

Payload mass looks to range from 0 to 6000 kg for the most part. Additionally, different launch locations appear to employ various payload masses.

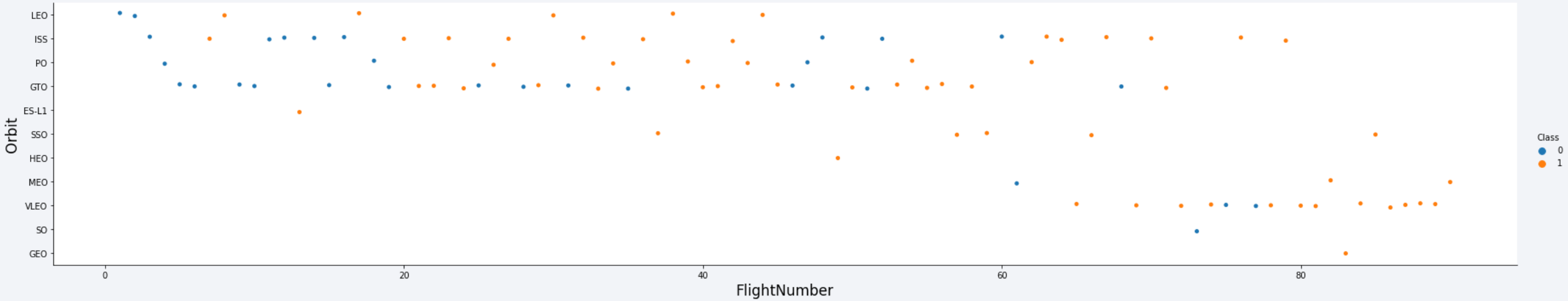
Success Rate vs. Orbit Type

Payload mass looks to range from 0 to 6000 kg for the most part.

- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- GTO (27) has the around 50% success rate but largest sample
- SO (1) has 0% success rate



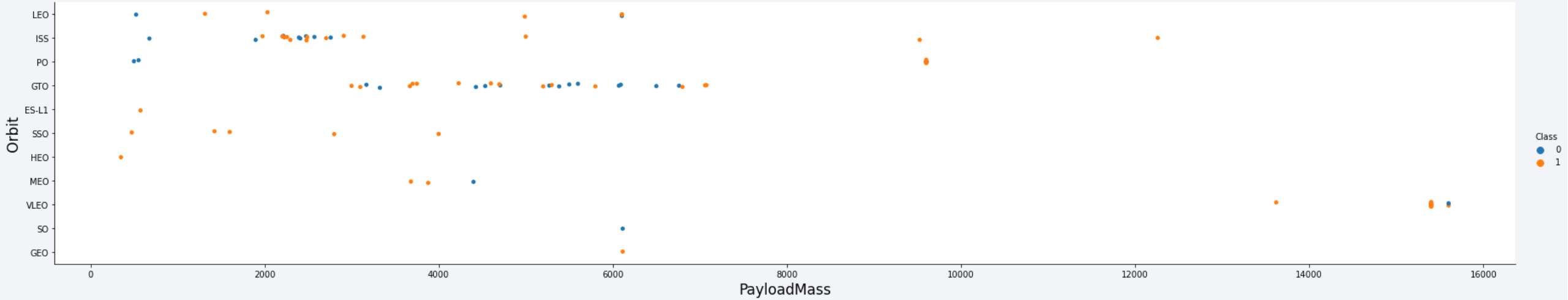
Flight Number vs. Orbit Type



Blue states unsuccessful launches while yellow states successful launch.

SpaceX began with LEO orbits, which had some success, then switched back to VLEO in more recent launches, suggesting that it performs better in lower orbits or orbits around the Sun.When choosing Launch Orbit instead of Flight Number, the Launch Outcome appears to be correlated with this decision.

Payload vs. Orbit Type

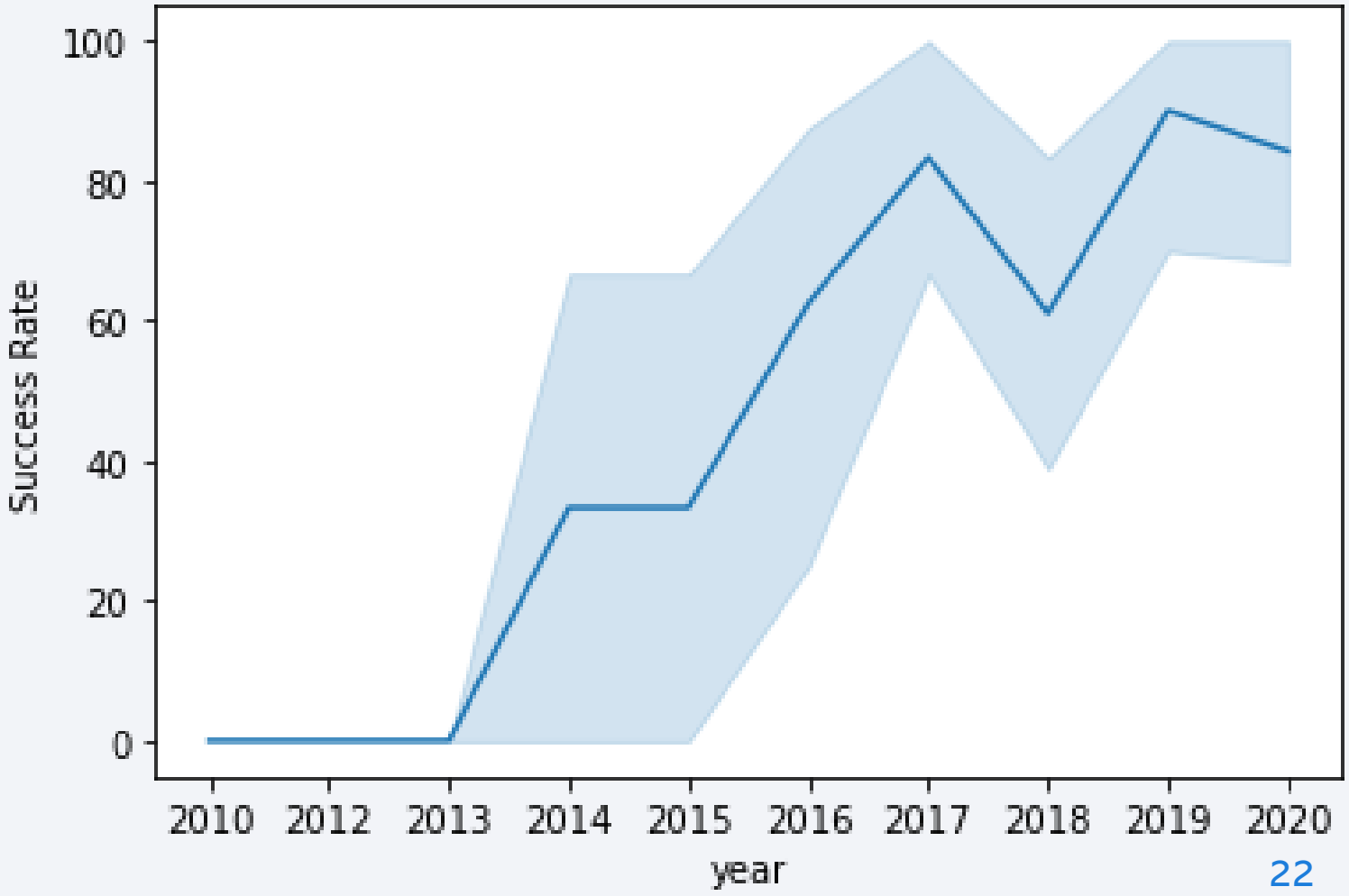


Blue states unsuccessful launches while yellow states successful launch.

The payload masses of LEO and SSO appear to be minimal. Only payload mass figures at the upper end of the range are available for the other most successful orbit VLEO. Thus, it appears that orbit and payload mass are proportional.

Launch Success Yearly Trend

- Since 2013, success has typically risen, with a tiny decline in 2018.
- Success rates in recent years have been around 80%.



All Launch Site Names

```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;
```

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

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

Both CCAFS SLC-40 and CCAFSSLC-40 most likely refer to the same launch location with incorrect data entry.

The former name might be CCAFS LC-40. Thus, probably only three distinct launch site values can be found:

VAFB SLC-4E, KSC LC-39A and CCAFS SLC-40.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD  LIKE '%CRS%';
```

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

TOTAL_PAYLOAD
111268

- CRS, or Commercial Resupply Services, is a designation for the International Space Station, where these payloads were delivered.
- When NASA was the client, this query adds up the total payload mass in kilogrammes.

Average Payload Mass by F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTBL WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVG_PAYLOAD
2928.4

The average payload mass for launches that utilised booster version F9 v1.1 is determined by this query.

F9 1.1's average payload mass seems to be on the lower end of our payload mass range.

First Successful Ground Landing Date

```
%sql SELECT min(DATE) AS DATE FROM SPACEXTBL WHERE MISSION_OUTCOME LIKE 'SUCCESS';
```

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

DATE
01-03-2013

The first successful ground pad landing date is returned by this query that seems to be in early 2013.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Mission_Outcome	QTY
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

This search yields a count of each mission result.

Nearly 99 percent of the time, SpaceX seems to succeed in achieving its mission objectives. This indicates that the majority of landing mishaps are deliberate.

Interestingly, one rocket's cargo status is unknown, and regrettably, one launch failed in flight.

Boosters Carried Maximum Payload

```
maxm = %sql select max(payload_mass__kg_) from SPACEXTBL
maxv = maxm[0][0]
%sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSIC
```

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

Booster_Version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- The results of this search are the booster iterations that could lift a payload of up to 15600 kg.
- These booster variants are all of the F9 B5 B10xx.x kind and are quite similar.

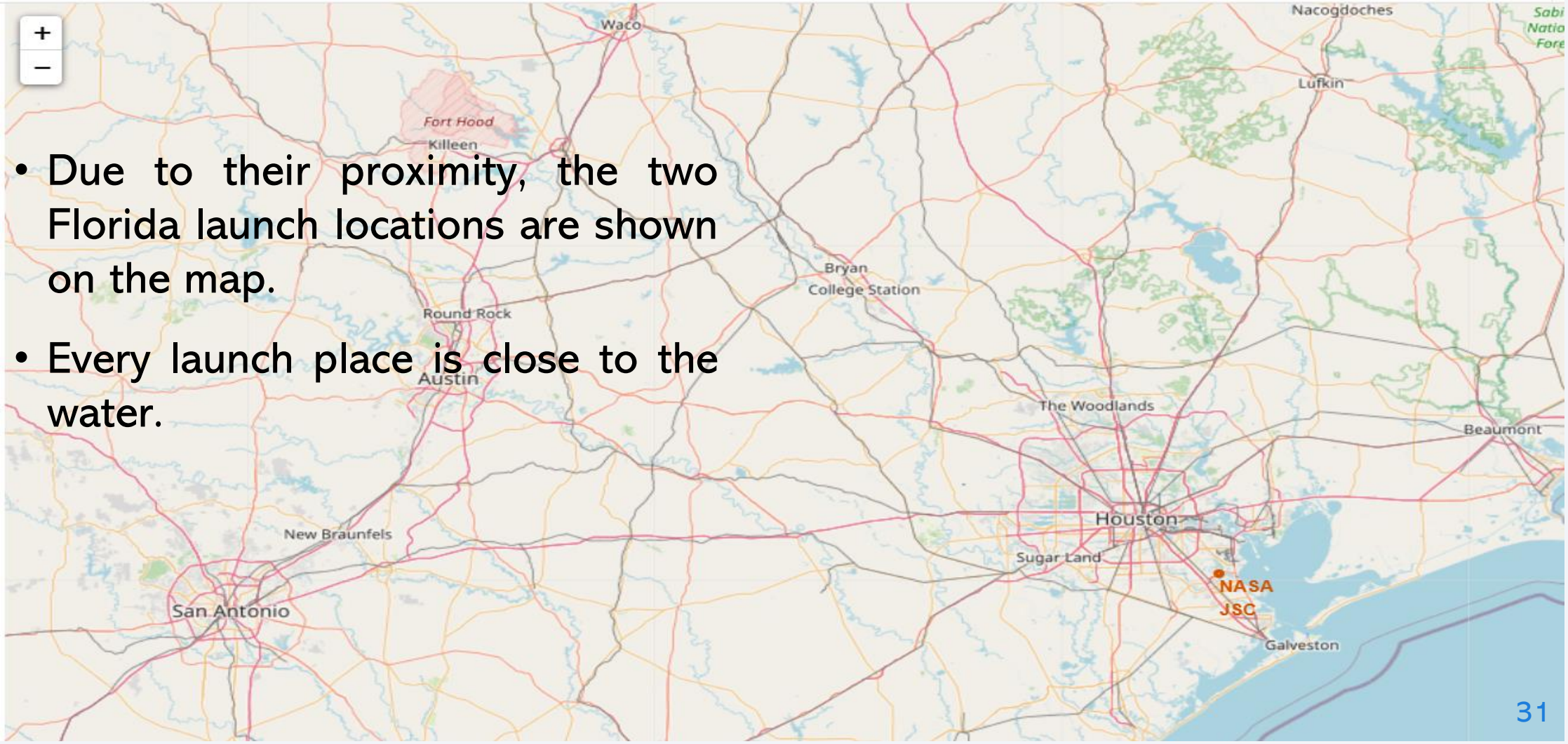
This suggests that the payload mass and the booster design are related.

Section 3

Launch Sites Proximities Analysis

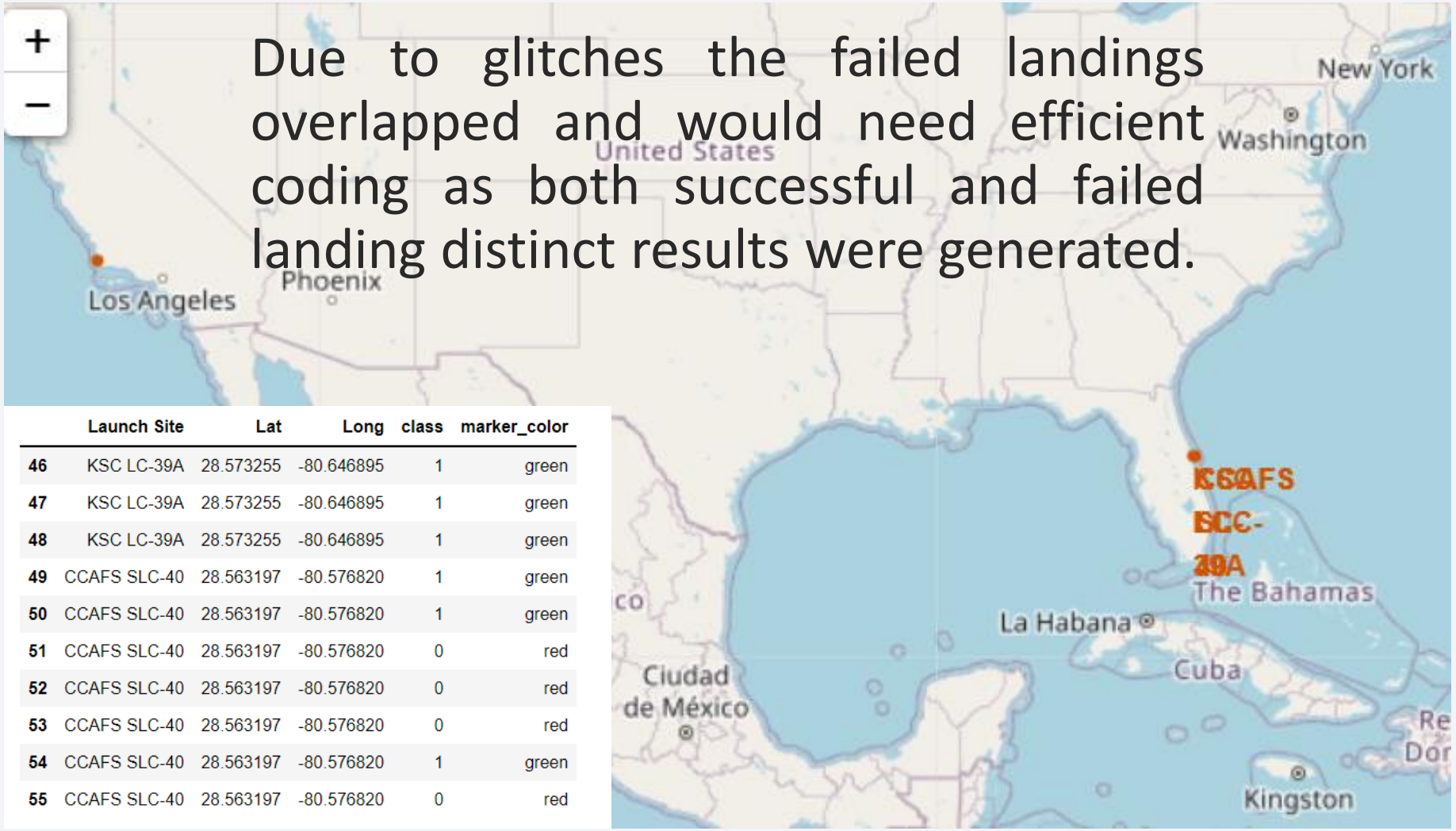
Launch Site Location

- Due to their proximity, the two Florida launch locations are shown on the map.
- Every launch place is close to the water.



Failed Launches for Each Site on the Map

Due to glitches the failed landings overlapped and would need efficient coding as both successful and failed landing distinct results were generated.

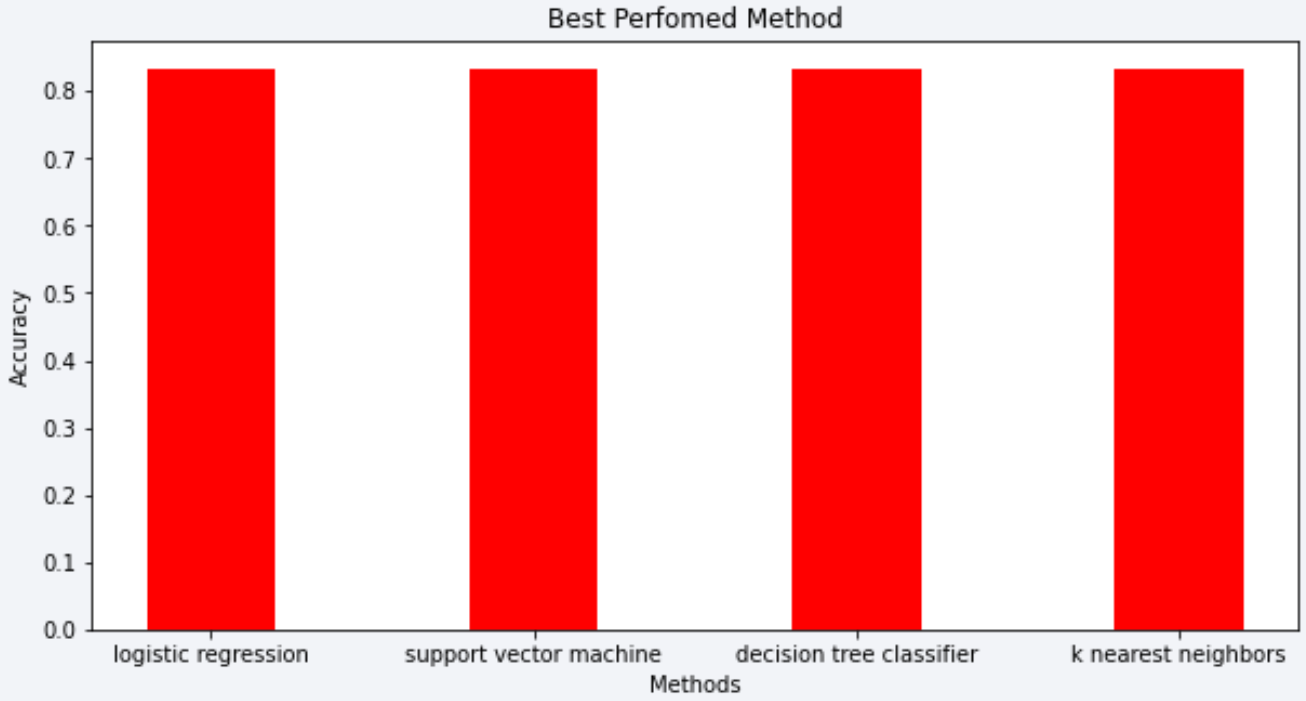


	Launch Site	Lat	Long	class	marker_color
46	KSC LC-39A	28.573255	-80.646895	1	green
47	KSC LC-39A	28.573255	-80.646895	1	green
48	KSC LC-39A	28.573255	-80.646895	1	green
49	CCAFS SLC-40	28.563197	-80.576820	1	green
50	CCAFS SLC-40	28.563197	-80.576820	1	green
51	CCAFS SLC-40	28.563197	-80.576820	0	red
52	CCAFS SLC-40	28.563197	-80.576820	0	red
53	CCAFS SLC-40	28.563197	-80.576820	0	red
54	CCAFS SLC-40	28.563197	-80.576820	1	green
55	CCAFS SLC-40	28.563197	-80.576820	0	red

Section 4

Predictive Analysis (Classification)

Classification Accuracy



All models were evaluated using a sample size of 18, which is relatively small for the building of an efficient learning model and almost equivalent accuracy on the test set at 83.33 percent

Thus, to choose one optimal model, we probably require additional information.

Confusion Matrix

The confusion matrix is the same for all models because their performance on the test set was identical.

- When successful landing was the genuine label, or True Positive, the models projected 12 successful landings.
- The models predicted three unsuccessful landings where the real label, or True Negative, was an unsuccessful landing
- .When the actual label was failed landings, the models projected three successful landings, which is known as a false positive.

Our prediction, thus, overstates the success of landings.

Conclusions

- The assignment was to create a machine learning model for Space Y, which is attempting to outbid SpaceX.
- 83 percent accurate machine learning model was produced.
- This model allows Allon Mask of SpaceY to forecast with a fair amount of accuracy if a launch will result in a successful landing before launch, allowing them to decide whether the launch should go forward or not.
- More information should ideally be gathered in order to choose the most accurate machine learning model.

Appendix

GitHub repository URL:

https://github.com/AJ-DarKnight/IBM_DataScience

Jupyter Notebooks on IBM individual links due to overuse of consumption LITE account:

- Data Collection: https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/fd03e6a8-4ae0-470e-ba62-6a40342496e0/view?access_token=4b0f18dfa12275b87d15d76a66cf2851b895e3a0f05239146f2a0931705bf36a
- EDA: https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/98fb065d-c8a1-4bbc-b6d9-4a72c9a9f69b/view?access_token=42e331aacdaf6253cc53d3623dec82c49ba79c6a65cb5e97f3876037afd86e4e
- EDA for Visualization: https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/c5898ee6-f46c-41d1-b055-3b7ea71ee90f/view?access_token=7be0ce40c9b547521830856c52a973ecd5b9de101b1d3e3eef47175b9bab3158
- Folium: https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/949bfe25-e11c-447c-a07d-463e76697b7c/view?access_token=e148fde438320993aa8ebd02ce0031b2fc264cad20a86df8285dac8120f5bc65
- ML: https://eu-gb.dataplatform.cloud.ibm.com/analytics/notebooks/v2/2bd0514f-21ff-4a45-96bc-ea3a6b2869ab/view?access_token=55f6eab7135c464344131ebfc23f5ae36f9721ddc1fae74081af174a338717f3