IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Christopher Charczuk
03/22/2022

Outline

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

Executive Summary

- Summary of methodologies
  - Collect Data
  - Data Wrangling
  - Exploratory Analysis
- Summary of all results
  - Visual Analytics Dashboard
  - Predictive Analysis

Introduction

A rival company, SpaceY, wants to compete with SpaceX and will need to determine the cost of each launch.

To help, we'll need to understand the factors that increase the chances of successful landings and reuse of rockets to help save costs.

Problems you want to find answers

- Which launch sites have the best chance for a successful landing?
- Does orbit help determine successful landings?
- Does payload help determine successful landings?
- Does Booster Version impact success of landings?

Section 1

Methodology

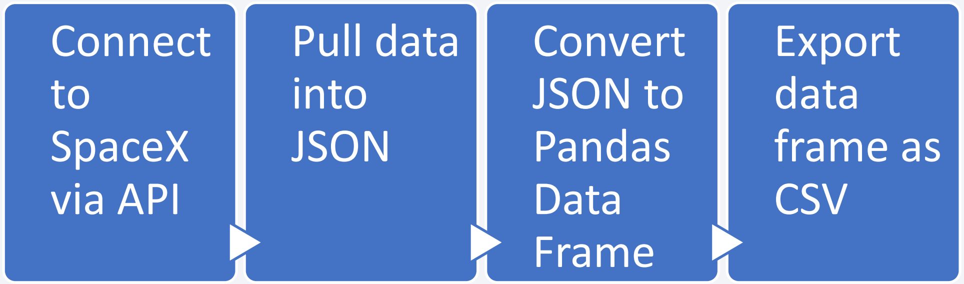
Methodology

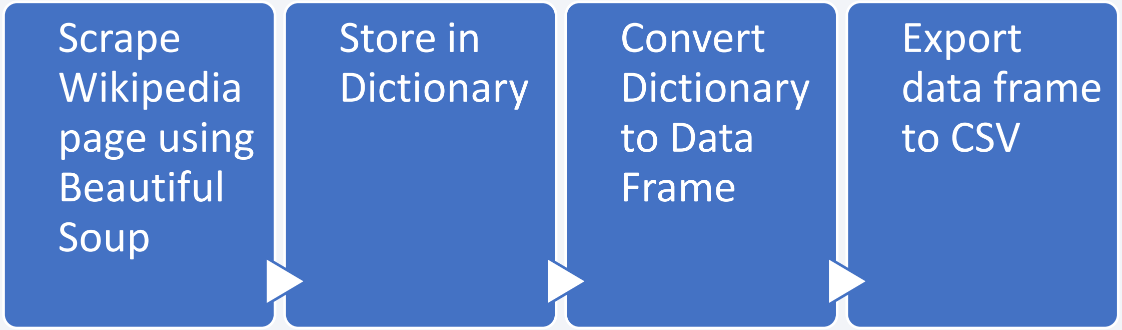
Executive Summary

- Data collection methodology:
    - The data was collected from SpaceX directly using an API. We also scraped a Wikipedia page for a table.
- Perform data wrangling
    - Created our dependent variable of landing success (1) or landing failure (0) and saved it with the Falcon 9 launch data.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Trained and tested 4 different models for classification predictions. The model with the most accuracy was chosen to predict successful landings of Falcon 9 rockets.

Data Collection

- How data sets were collected:
  - Connected to SpaceX data via API and pulled required data into JSON
    - Converted JSON into a Pandas data frame and export/save as a CSV.
  - Scrape Wikipedia page using Beautiful Soup and store in dictionary.
    - Convert dictionary to data frame in Pandas and export/save as CSV.





Data Collection – SpaceX API

Assign URL and Pull Data

- #assign URL and pull data
- spacex_url="https://api.spacexdata.com/v4/launches/past"
- response = requests.get(spacex_url)

Static response JSON object

- #Use assigned static response JSON object
- static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-
- DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'

Normalize as Data Frame

- #Use json_normalize method to convert the json result into a dataframe
- df = pd.json_normalize(requests.get(static_json_url).json())

Clean Data

- Not shown here (code too long)

Save Falcon 9 Data as CSV

- #save Falcon 9 data as csv
- data_falcon9.to_csv('dataset_part_1.csv', index=False)

GitHub File

Data Collection - Scraping

Scrape Wikipedia

- static_url = "https://en.wikipedia.org/w/index.php?title=List_of_
- Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
- page=requests.get(static_url)

Store as Beautiful Soup HTML object

- soup=BeautifulSoup(page.text,'html.parser')

Parse and Convert to Data Frame

- Create dictionary with column names from data frame in last slide.
- df=pd.DataFrame(launch_dict)

Save as CSV

- df.to_csv('spacex_web_scraped.csv', index=False)

GitHub File

Data Wrangling

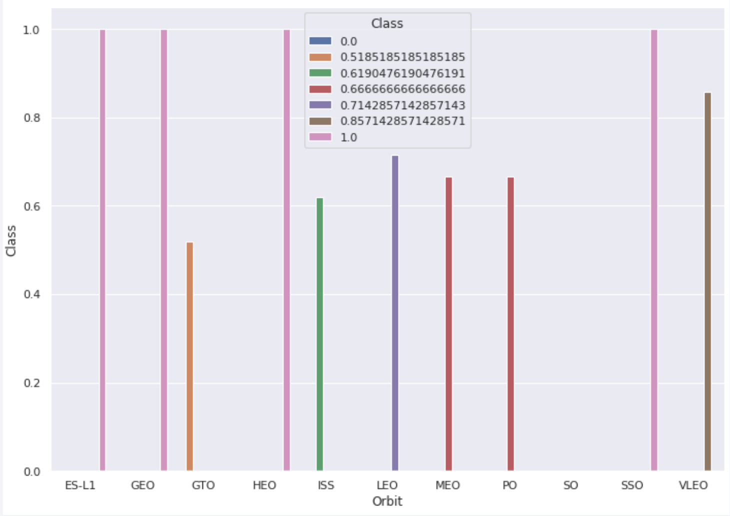
- Loaded data set containing Falcon X launches.
- Created a dependent variable named Class and added it to our data frame:
  - 1 = successful landing
  - 0 = unsuccessful landing

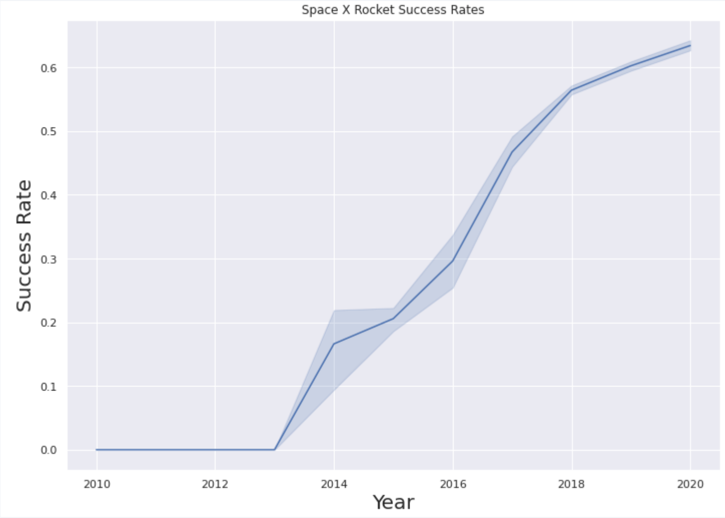
Load CSV containing Falcon X data → Create a key for bad outcomes → Create vector named Class to contain bad outcomes (0) or successes (1) → Add Class vector to loaded data frame → Save as CSV

[GitHub File]

EDA with Data Visualization

- Success rate by orbit helps to identify orbits with chance for higher successes.
- Success rate over time shows that successes are improving each year and that diminishing returns seems to be around 65%-70%





[GitHub File]

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- Rank 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

GitHub File

Build an Interactive Map with Folium

- Added a marker for each launch site
- Added success/failed landings for each launch site marker
- Calculated the distance between launch sites and other map locations

The markers show where each launch site is located on the map. The success/failed attempts for each launch site show when each site is clicked and zoomed-in on.

GitHub File

Build a Dashboard with Plotly Dash

- Pie chart to show total success percent for all locations.
- The pie chart will show success/failure percentage for an individual location when selected from the drop down.
- Payload and success plot shows the successes and failures for the ranges of payloads and by booster version.

GitHub File

Predictive Analysis (Classification)

- Defined our dependent variable (y = Class column)
- Defined our X variables as a data frame and preprocessed/scaled the data.
- Split the dependent variable and independent variables in to a training set (80%) and a test set (20%)
- Train models using Logistic Regression, Support Vector Machine (SVM), Decision Tree, and k Nearest Neighbors (kNN).
- Run test data to measure accuracy of models and select the best model.
- Add the GitHub URL of your completed predictive analysis lab, as an external reference and peer-review purpose

Define and scale y → Define and scale IVs (X) → Split into training and test sets → Create models using 4 different approaches to ML → Run Test set through models → Select best performing model based on Test set performance.

GitHub Link

Results

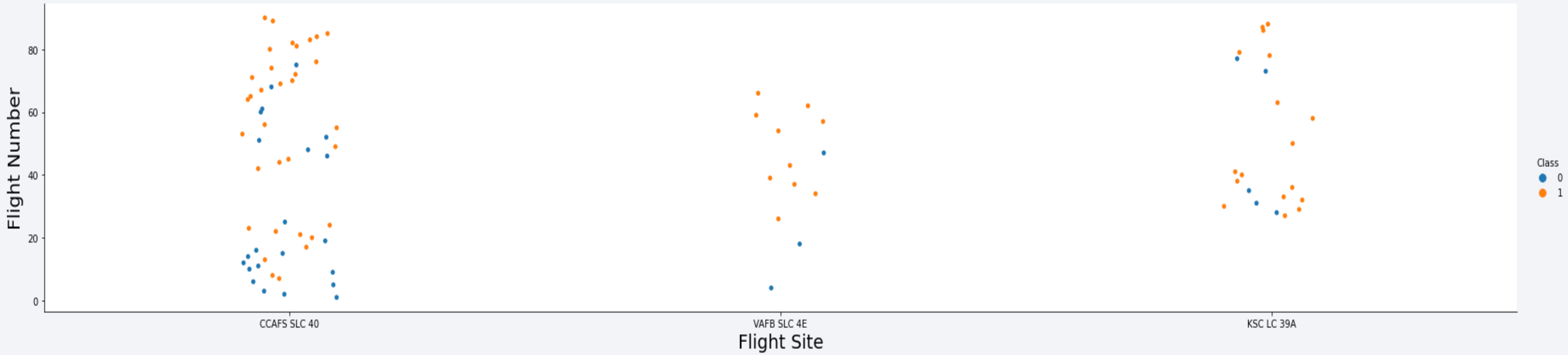
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

- CCAFS SLC 40 has the most launches and more successes as the flight number increases.
- WAFB SLC 4E and KSC LC 39A have fewer launches, but a higher percentage of successes.

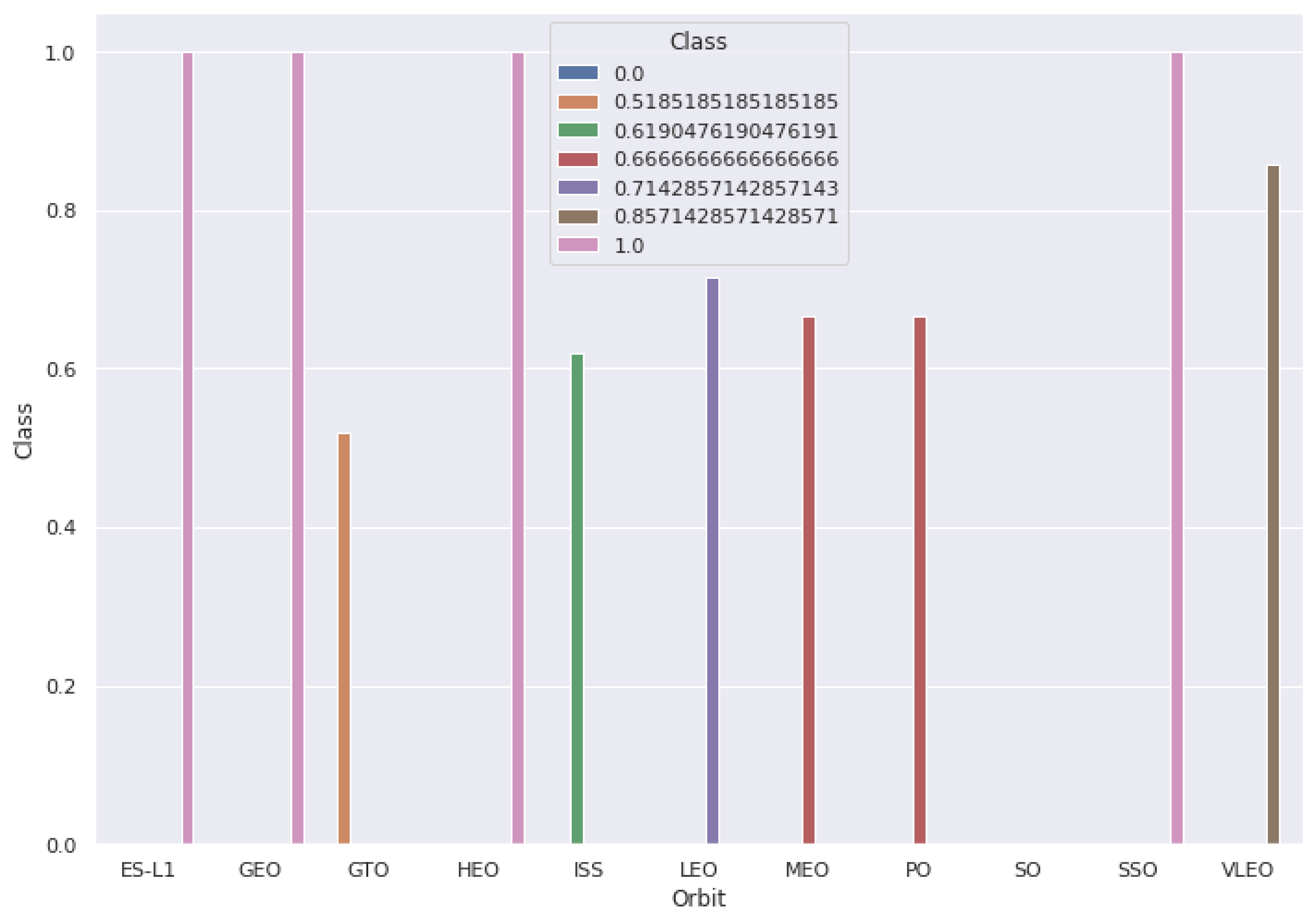


Payload vs. Launch Site

- Success rate is lower for low payloads at SLC 40. But high pay loads have a much higher number of successes.
- There's no clear pattern to successes by pay load for SLC 4E or LC 39A. Successes run the gamut of pay load weight.

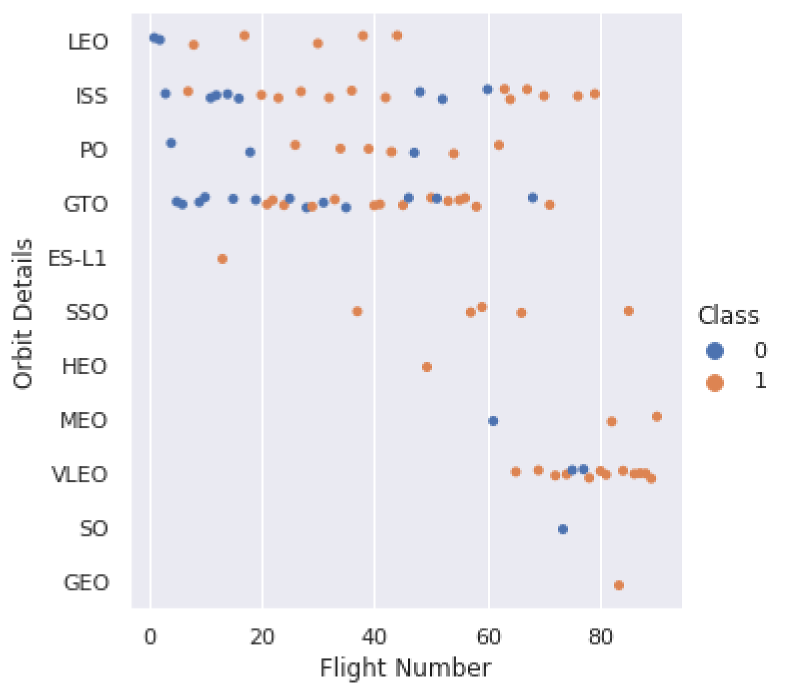
Success Rate vs. Orbit Type

- 100% success rate for ES-L1, GEO, HEO, and SSO.
- Success rates lower than 2/3rds (66%) for GTO, ISS, MEO, and PO.



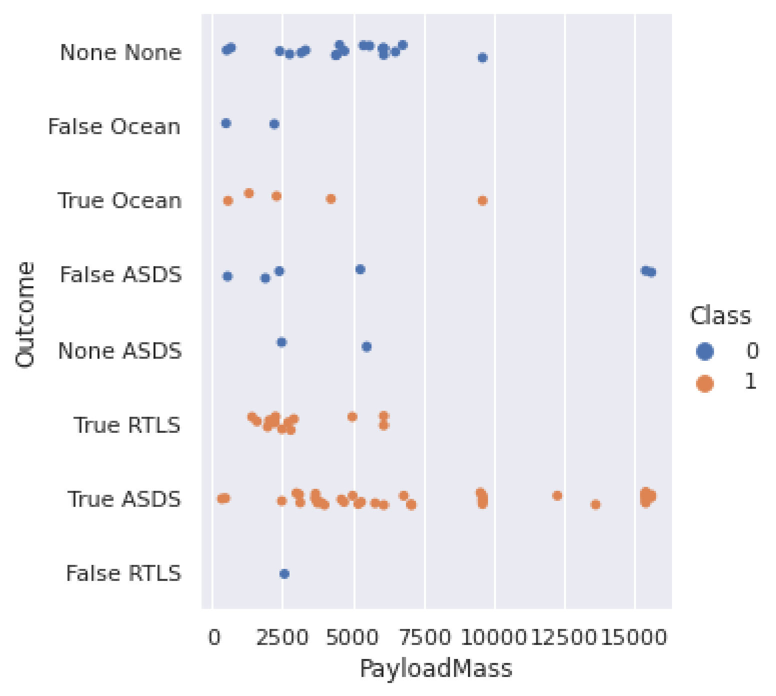
Flight Number vs. Orbit Type

- There's a general trend of more successes as flight number increases (regardless of orbit).
- Orbits with 100% success rate have low sample sizes (<4).



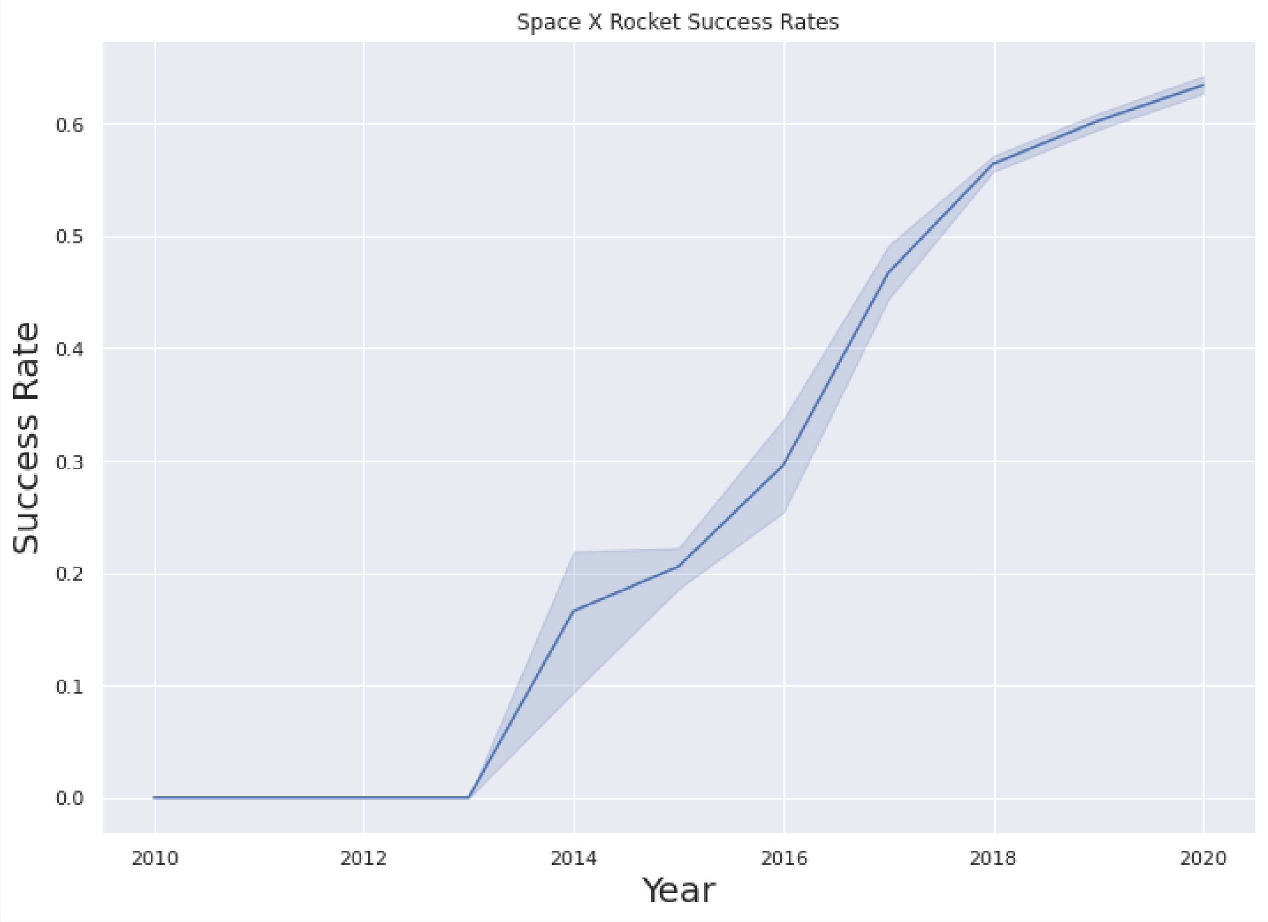
Payload vs. Orbit Type

- Many of the unsuccessful landings have no defined outcome.
- There are far fewer landing failures if we remove Outcomes with no orbit class or True/False indicator.



Launch Success Yearly Trend

- Success rate improves as time progresses.
- There appears to be the start of a diminishing returns curve around 60%.
- This indicates that it may be more challenging to improve success rate from 65% up to 100%.



All Launch Site Names

- Used distinct to only list each site once.

```
%sql select distinct launch_site from SPACEXDATASET;
```

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

Launch Site Names Begin with 'CCA'

- Displayed first 5 rows where launch site name started with “CCA”

```
%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
```

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 payload carried by boosters from NASA: 45,596.

```
%sql select sum(payload_mass__kg_) as payload from SPACEXDATASET where customer = 'NASA (CRS)';
```

payload
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1 = 2,534

```
%sql select avg(payload_mass__kg_) as payload_avg from SPACEXDATASET where booster_version like 'F9 v1.1%';
```

payload_avg
2534

First Successful Ground Landing Date

- Date of the first successful landing was 12/22/2015
- Found the minimum date where landing outcome was on a ground pad and a success.

```
%sql select min(DATE) as earliest_success from SPACEXDATASET where lower (landing__outcome) like '%ground%' and lower(landing__o
utcome) like '%success%';
```

earliest_success
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Used distinct to list each booster with successful drone ship landings.
- Use where clause to narrow down to correct pay load range and landing outcome (drone ship).

```
%sql select distinct booster_version from SPACEXDATASET where payload_mass__kg_ between 4000 and 6000 and lower(landing__outcom
e) like '%drone ship%';
```

booster_version
F9 FT B1021.2
F9 FT B1031.2
F9 FT B1020
F9 FT B1022
F9 FT B1026

Total Number of Successful and Failure Mission Outcomes

- There were 99 successful outcomes and 1 failures.

```
%sql select count(mission_outcome) from SPACEXDATASET where mission_outcome = 'Success';
```

1
99

Boosters Carried Maximum Payload

- I was surprised to see multiple booster versions that all had the max, so the second column was added to verify.

```
%sql select distinct booster_version, max(payload_mass__kg_) as max_payload from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET) group by booster_version;
```

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

2015 Launch Records

- There were two failures on the drone ship.

```
%sql select booster_version, launch_site, DATE, landing__outcome from SPACEXDATASET where year(DATE) = 2015 and lower(landing__o
utcome) like '%failure%' and landing__outcome like '%drone ship%';
```

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

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

- Pulled count of landing outcomes and ordered them in descending order by highest count to lowest.

```
%sql select distinct landing__outcome, count(landing__outcome) as number from SPACEXDATASET where DATE between '2012-06-04' and '2017-03-20' group by landing__outcome order by number desc;
```

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

Section 3

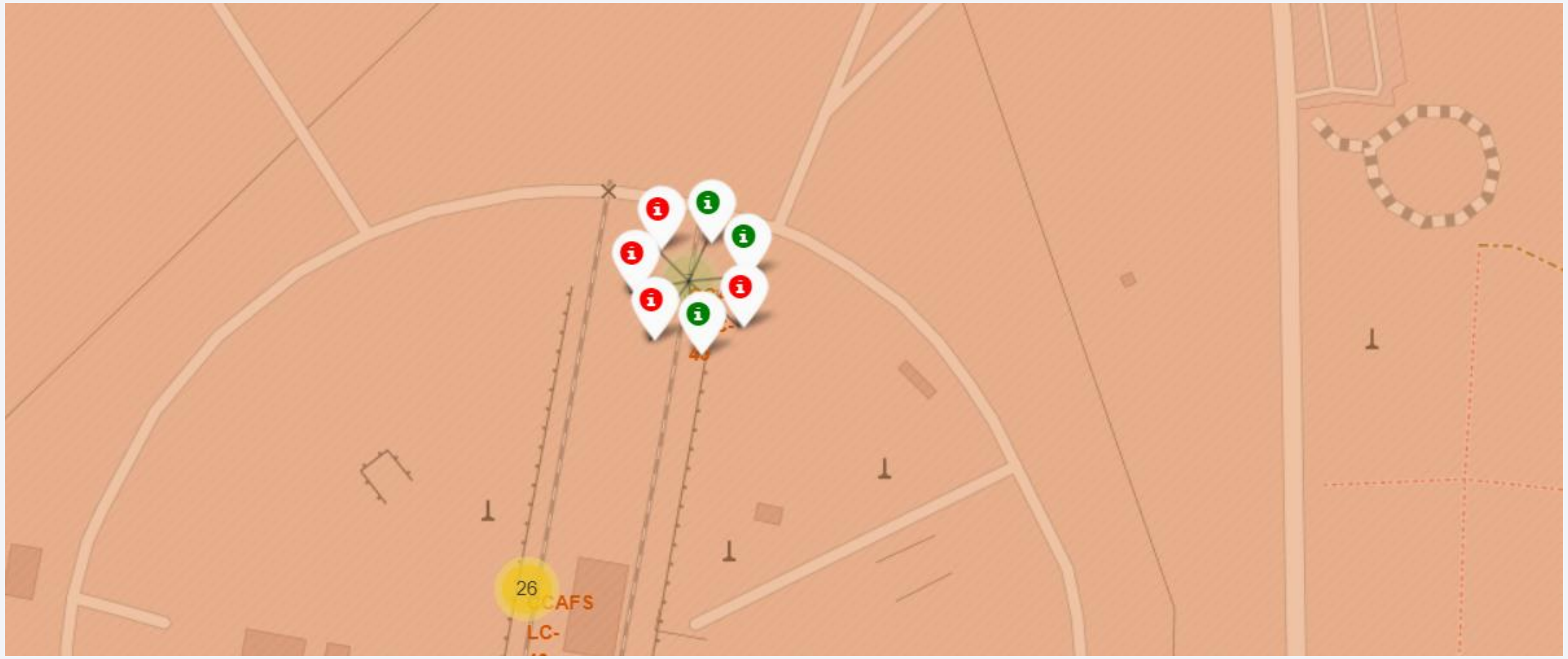
Launch Sites Proximities Analysis

Falcon 9 Launch Sites

All launch sites for Falcon 9 are near the coast and in California or Florida.

Landing Success from Launches at CCAFS SLC-40

- CCAFS SLC-40 has more failures than successes.



CCAFS SLC-40 Distance to Atlantic Ocean

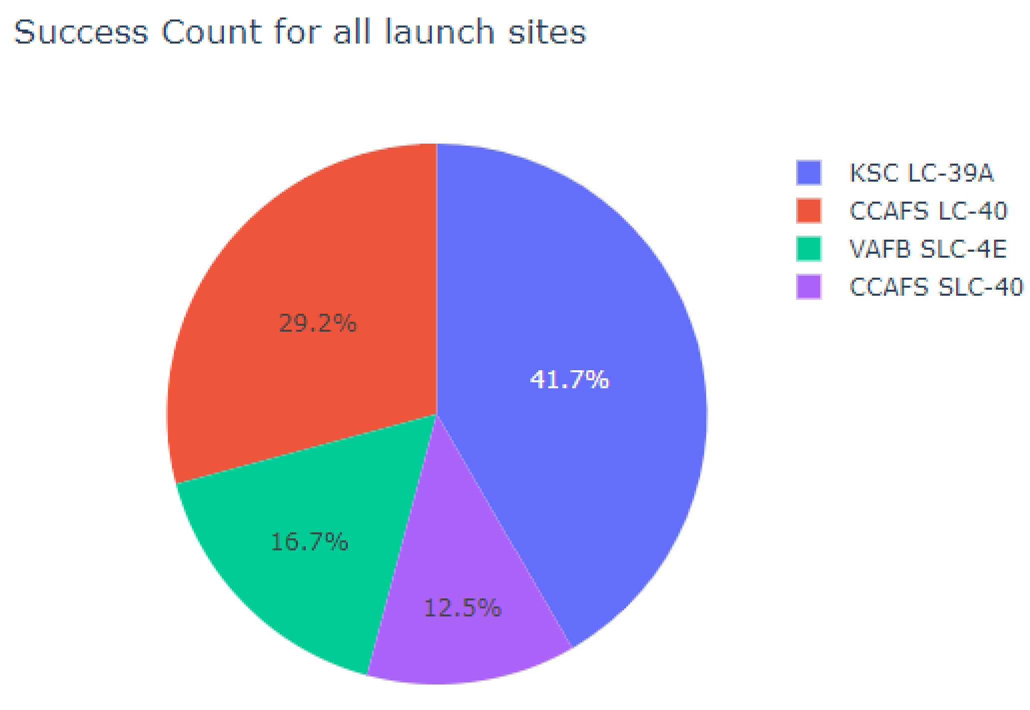
- CCAFS SLC-40 is 0.87 KM from the Atlantic Ocean.

Section 4

Build a Dashboard with Plotly Dash

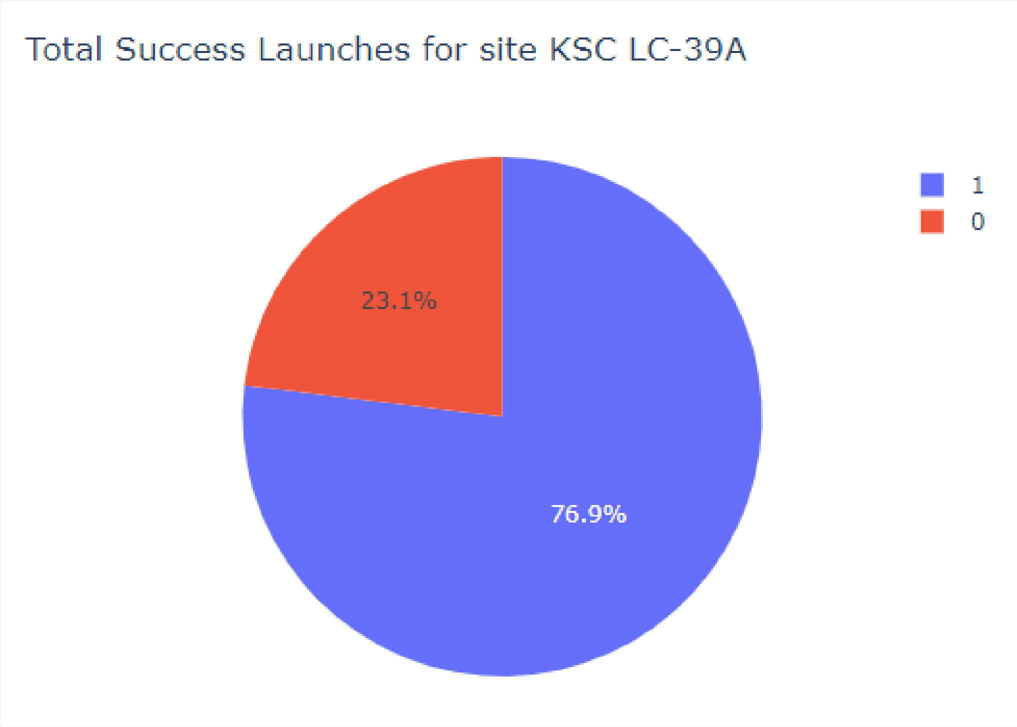
Highest Success by Launch Site

- KSC LC-39A has the highest percentage of total successes (41.7%).
- CCAFS SLC-40 has the lowest total successes (12.5%).



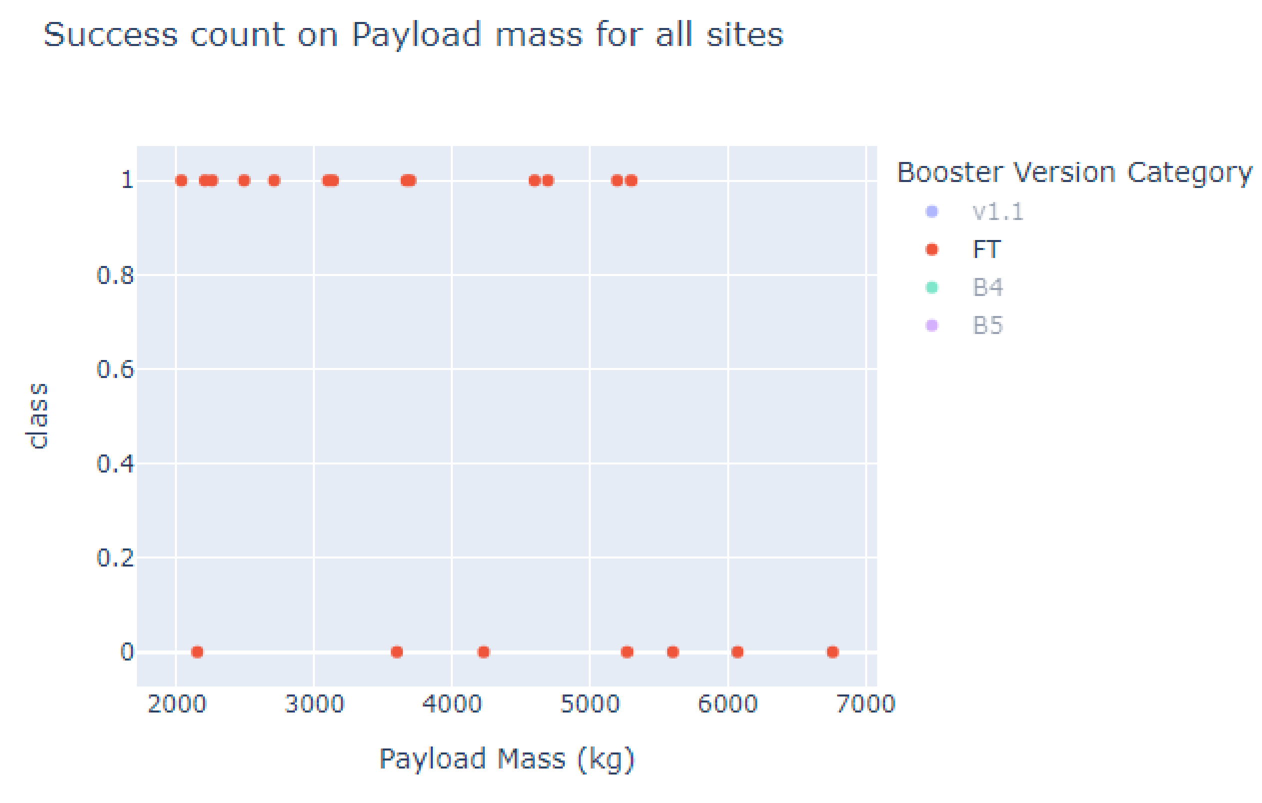
KSC LC-39A Success Rate

- 76.9% of launches from KSC LC-39A land successfully.



FT Leads Booster Version with Total Successes

- For payloads between 2000kg and 7000kg, the FT booster version has the highest number of successes.

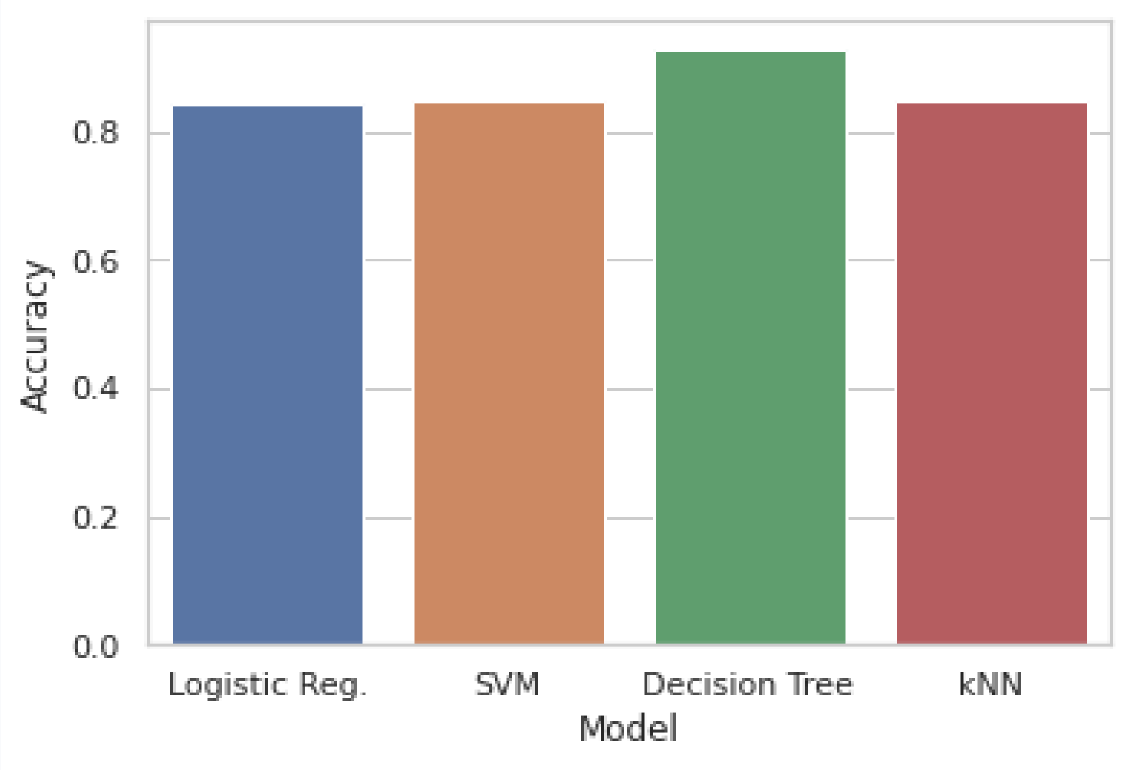


Section 5

Predictive Analysis (Classification)

Classification Accuracy

- Decision Tree has the highest accuracy for predicting successful landings at 92.86%.



Confusion Matrix

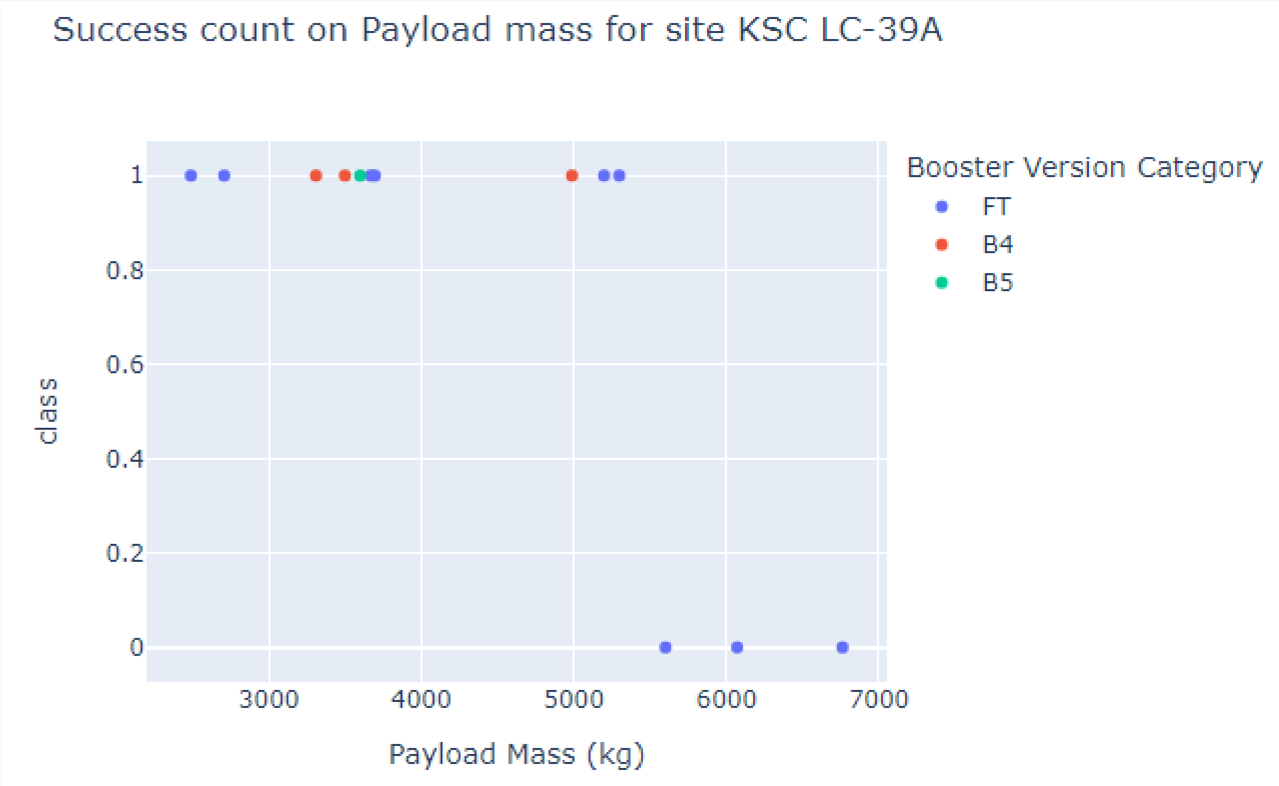
- The error in the Decision Tree model was predicting false positive landings (3).
- It predicted all the successful landings accurately.
- It predicted half of the unsuccessful landings.

Conclusions

- Successful landings have increased over the years. Each year continues to have higher success rates.
  - Earlier booster versions had very poor landing success rates. It's possible that improved boosters over time are improving performance (appendix).
- It's better to launch from VAFB SLC 4E and KSC LC 39A than from CCAFS SLC 40 because they have higher mission success rates.
  - They also had fewer v1.0 and v1.1 booster version launches.
- Decision Tree modeling was the best at predicting successful landings for Falcon 9 (92.86%).
- There's no clarity on orbit being important since the orbits with the highest success rates had the fewest attempts.

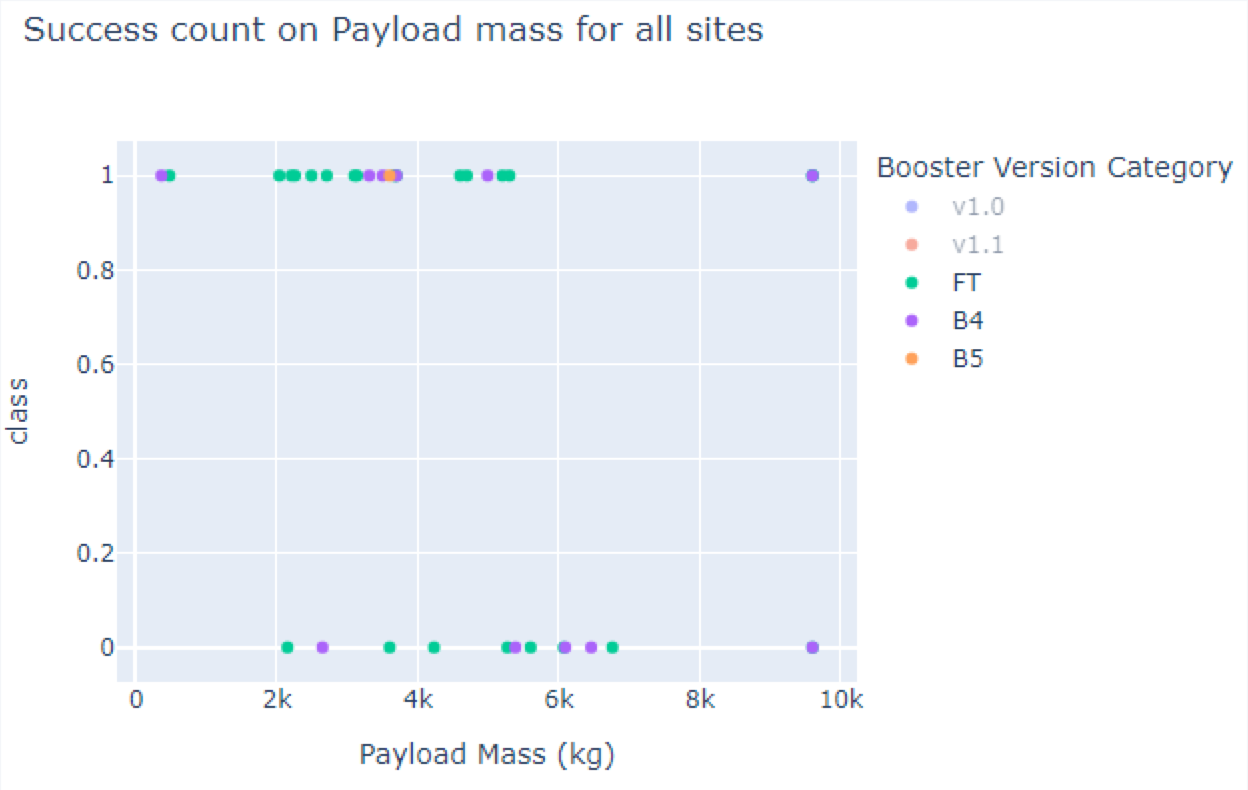
Appendix

KSC LC-39A has a high total of successful landings because the most successful Booster Versions are used from this location.



Appendix

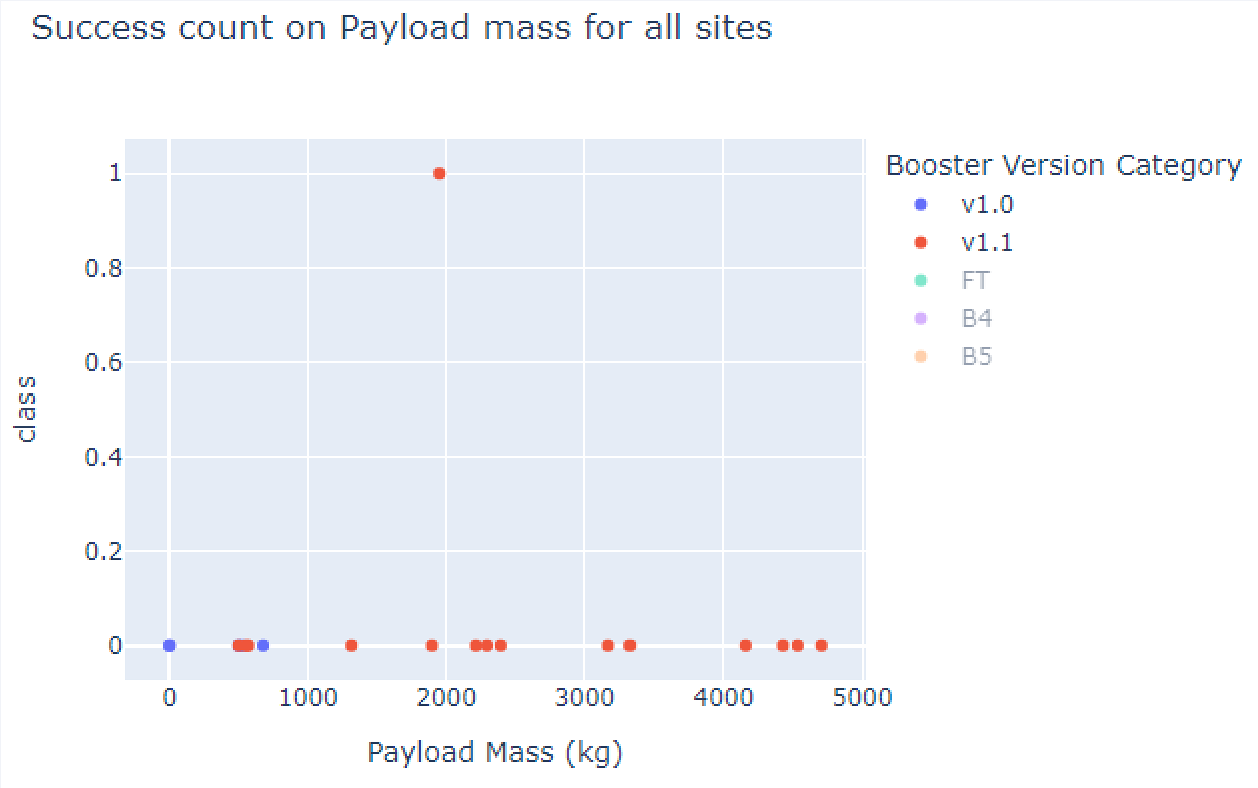
- FT, B4, and B5 Booster Versions have more successes than failures.



Appendix

- V1.0 and v1.1 have only one successful landing.

- If these two were used early in testing, this could explain why successes have improved over time: new boosters are helping to increase the chance for a successful landing.



Thank you!