SpaceX Launch Analysis

Kevin Benson

Feb. 17th, 2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - SQL
  - Folium
  - Dashboard
  - Predictive Analysis
- Conclusion
- Appendix

EXECUTIVE SUMMARY

Data was obtained, cleaned, explored, then analyzed from SpaceX. This data was then used to determine if a launch will be successful or not.

Visualization with plotly and folium aided in getting a grasp of the relation between variables in this dataset, in addition to SQL commands.

Several machine learning models provided through Scikitlearn were tested to predict the outcome of a launch. Logistic regression, SVM, Decision Tree and KNN were all the best at predicting the launch outcome.

INTRODUCTION

It was set out to determine if the first stage of a SpaceX rocket will successfully land, and therefore be reused. This drastically lowers the cost of launching rockets.

SpaceX provides launch data on their website for free. This data was refined and analyzed using several methods detailed in this report.

If we can successfully predict a safe launch, then this is advantageous and will save the company lots of money.

METHODOLOGY: Data Collection/Wrangling

Data was collected from SpaceX using an API (https://api.spacexdata.com) in addition to web scraping with BeautifulSoup. Then, data was wrangled into a pandas dataframe with a 'Class' column indicating a launches result.

Core data was collected with functions doing API calls with the python requests module. Some of the core data column titles were: *BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Longitude, and Latitude.*

The webscraping portion of the data collection phase was done using an HTML table from wikipedia. Once the data had been collected, it was wrangled. Launch results were interpreted into '0' or '1' in the 'Class' column. This was then used as the independent variable in predictive analysis.

METHODOLOGY: EDA/Visual Analytics

Exploratory Data Analysis (EDA) was completed using SQL and visual analysis was done with folium, matplotlib, and seaborn. Using these graphical methods, the relationships between dependent variables and 'Class' was better understood.

For the SQL portion of EDA a table was uploaded to IBM's cloud server, then queries were performed with help from the sqalchemy toolkit for python. Queries were performed to extract information such as unique launch sites, average payload masses, different landing outcomes.

An interactive dashboard was made with plotly and dash to even further explore the data. The dashboard contains a dropdown menu to switch between landing sites and a slider to select a desired range of payload masses. Callback functions were used to build these interactive components.

METHODOLOGY: Predictive Analysis

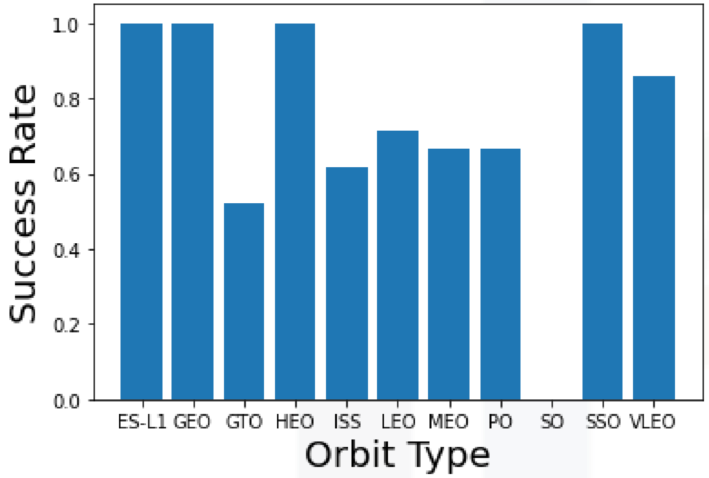
Lastly, several machine learning models were fit to training data extracted from the dataset and scored on their abilities to successfully predict the outcome of a launch from test data.

Using the sklearn python library, the data was pre-processed by standardizing with *StandardScaler*, then split 80/20 into training and testing sets with *train_test_split.*

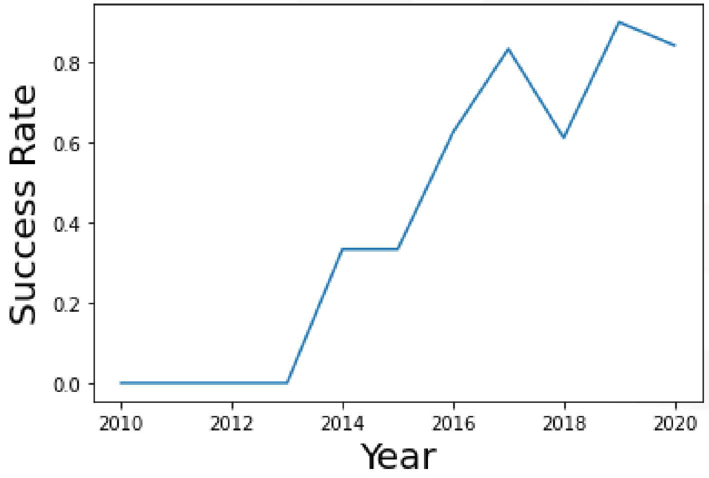
A *GridSearchCV* object was created for sever machine learning models, then fit to the training data. The models in question were: logistic regression, SVM, decision tree, and KNN.

The performance of these models was then analyzed via the R-squared score and by plotting a confusion matrix.

RESULTS: EDA with Visual Analysis



- Some orbit types are indicative of a higher success rate
- ES-L1, GEO, HEO, and SSO all has a success rate of 100%
- GTO was the least successful orbit type at 50%



- There was not a success for the years 2010-2013
- From 2013-2017 the success rate increased from 0% to 80%
- The success rate decreased from 2017 to 2018, and again from 2019 to 2020

see Figures A & B in the Appendix for scatterplots

Results: EDA with SQL (Tasks 1-4)

Display the names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT launch_site from spacex_data
```

* sqlite:///spacex.db
Done.

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

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%%sql
SELECT customer, sum(payload_mass_kg) as 'Total Payload Mass' from spacex_data
WHERE customer
    LIKE 'NASA (CRS)'
```

* sqlite:///spacex.db
Done.

Customer	Total Payload Mass
NASA (CRS)	45596

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%%sql
SELECT * from spacex_data
WHERE launch_site
    LIKE 'CCA%'
LIMIT 5
```

* sqlite:///spacex.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

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

```
%%sql
SELECT booster_version, AVG(payload_mass_kg) as 'Average Payload Mass' from spacex_data
WHERE booster_version
    LIKE 'F9 v1.1%'
```

* sqlite:///spacex.db
Done.

Booster_Version	Average Payload Mass
F9 v1.1 B1003	2534.6666666666665

Results: EDA with SQL (Tasks 6-8)

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
%%sql
SELECT MIN(date) as 'First Success' from spacex_data
WHERE landing_outcome
    LIKE 'Success (ground pad)'
```

* sqlite:///spacex.db
Done.

First Success
01-05-2017

List the total number of successful and failure mission outcomes

```
%%sql
SELECT count(*) as 'Number of Successful' from spacex_data
WHERE mission_outcome LIKE 'Success%'
```

* sqlite:///spacex.db
Done.

Number of Successful
100

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%%sql
SELECT booster_version, payload from spacex_data
WHERE landing_outcome
    LIKE 'Success (drone ship)'
    AND payload_mass_kg > 4000
    AND payload_mass_kg < 6000
```

* sqlite:///spacex.db
Done.

Booster_Version	Payload
F9 FT B1022	JCSAT-14
F9 FT B1026	JCSAT-16
F9 FT B1021.2	SES-10
F9 FT B1031.2	SES-11 / EchoStar 105

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%%sql
SELECT booster_version from spacex_data
WHERE payload_mass_kg =
    (SELECT max(payload_mass_kg) from spacex_data)
```

* sqlite:///spacex.db
Done.

Booster_Version
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

IBM Developer

Results: EDA with SQL (Tasks 9-10)

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

```
%%sql
SELECT landing_outcome, booster_version, launch_site from spacex_data
WHERE landing_outcome
    LIKE 'Failure (drone ship)'
    AND date
    LIKE '%2015%'
```

 * sqlite:///spacex.db
Done.

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

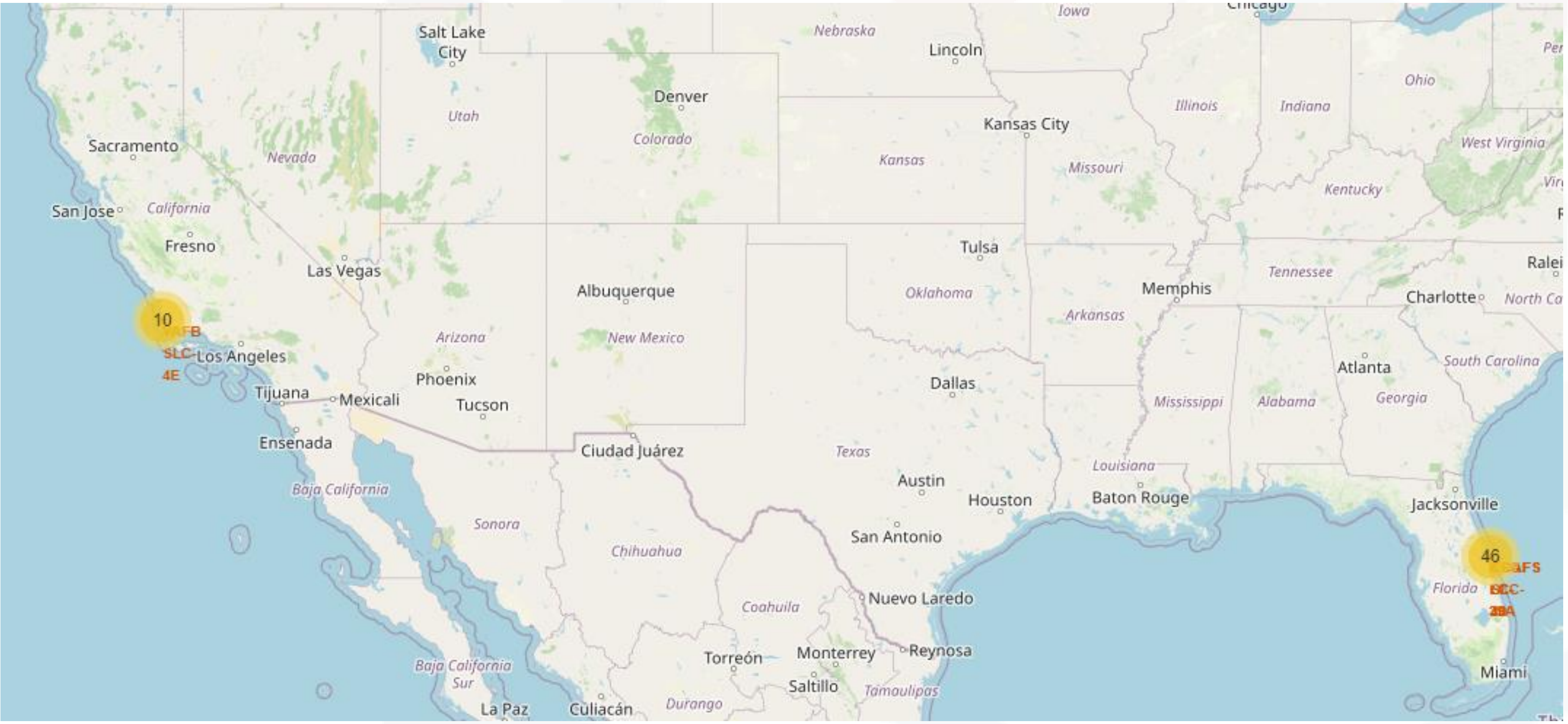
Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%%sql
SELECT landing_outcome, count(landing_outcome) from spacex_data
WHERE date
    BETWEEN '06-04-2010' AND '20-03-2017'
GROUP BY landing_outcome
ORDER BY count(landing_outcome) desc
```

 * sqlite:///spacex.db
Done.

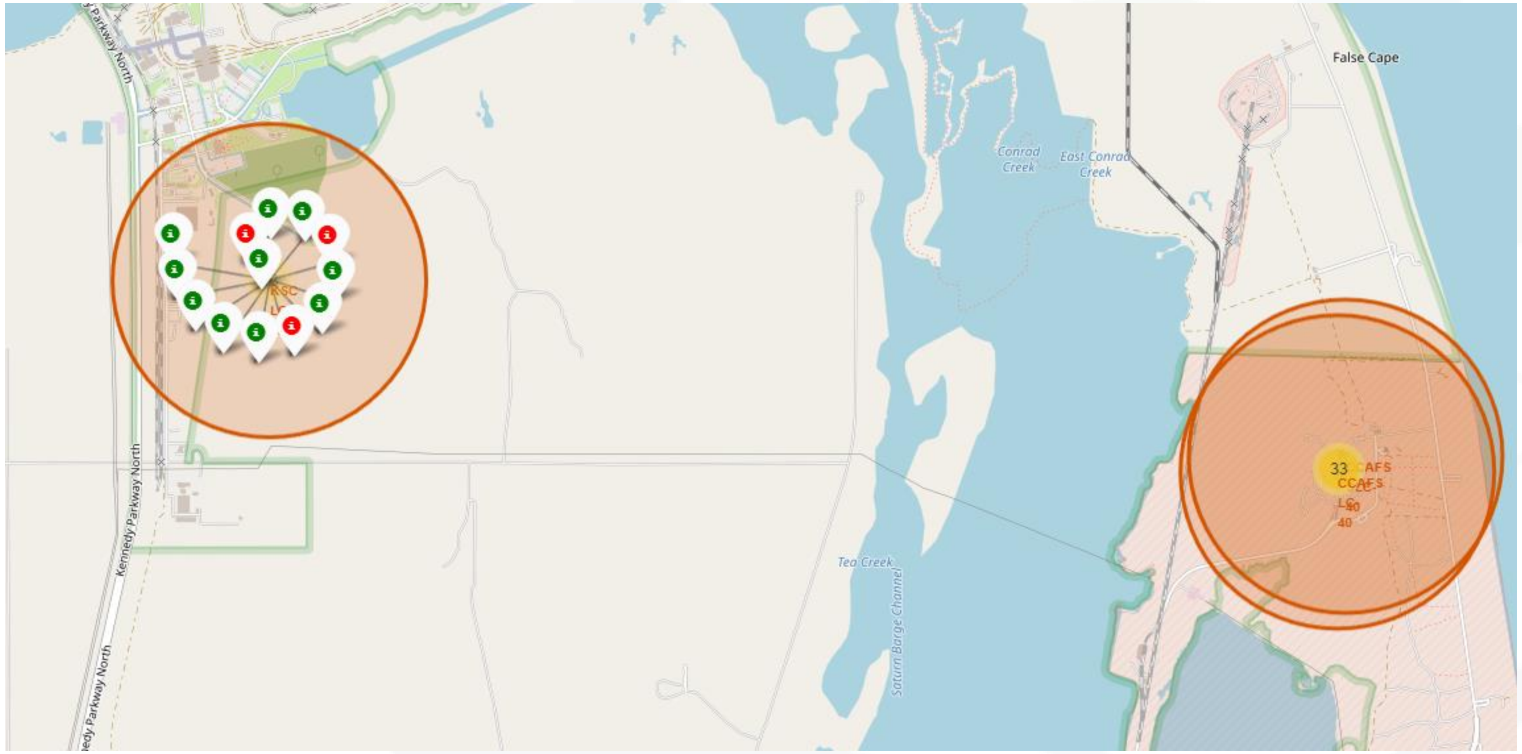
Landing_Outcome	count(landing_outcome)
Success	17
Success (drone ship)	8
Success (ground pad)	6
No attempt	5
Failure (drone ship)	4
Controlled (ocean)	3
Failure	2
No attempt	1
Failure (parachute)	1

Interactive Folium Map



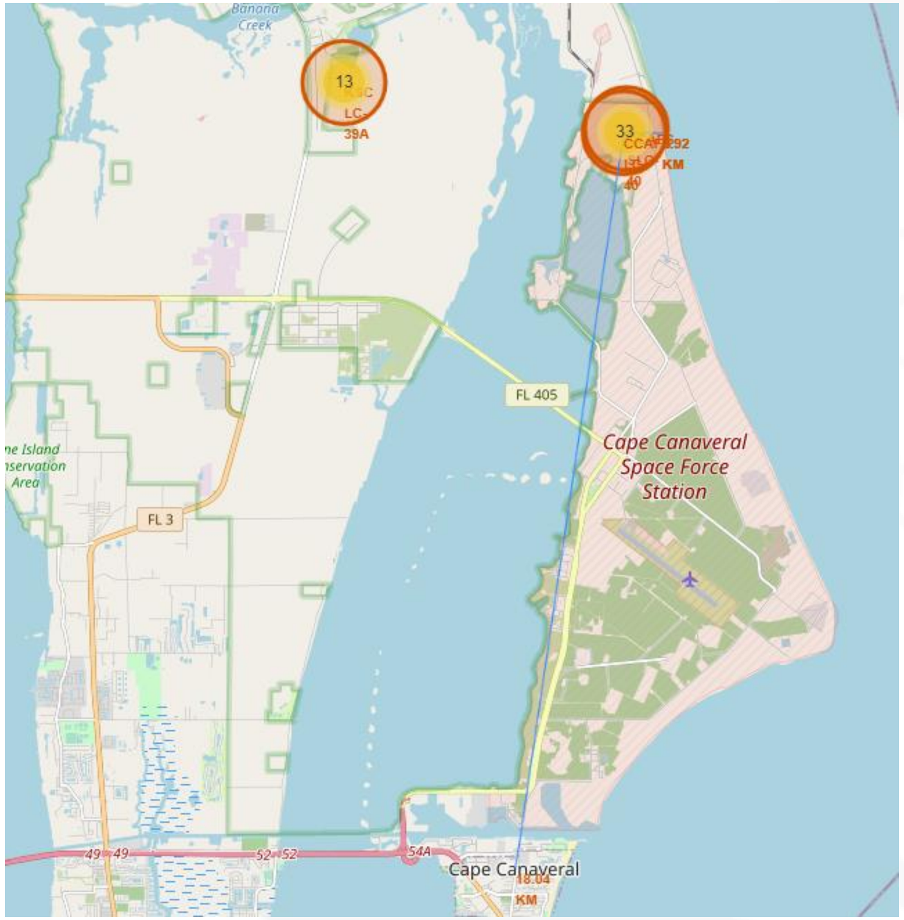
- VAFB SLC-4E is the only launch site on the west coast. All three other launch sites are in Florida
- 10 out of 56 launches have occurred at VABFC SLC-4E

Interactive Folium Map



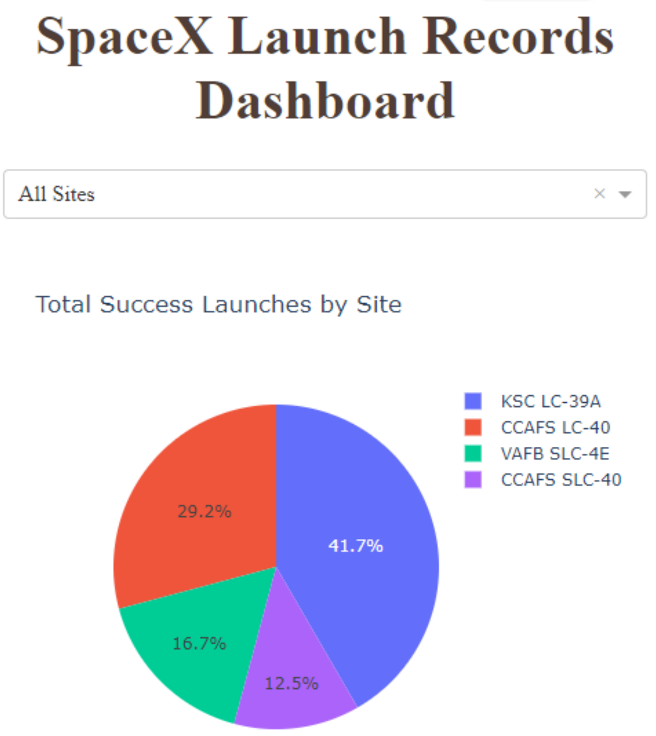
- Three of the launch sites are on the coast of Florida, near Cape Canaveral
- Two of them are right on top of each other
- Three out of thirteen launches at KSC LC-39A were unsuccessful

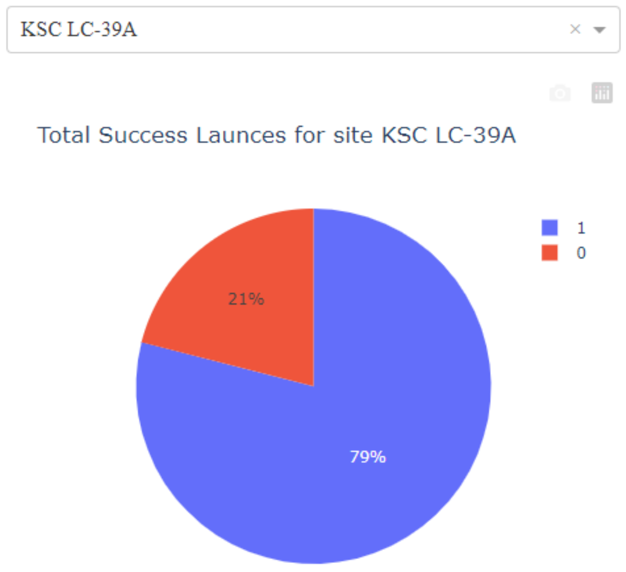
Interactive Folium Map



- The distance from CCAFS LC-40 to Cape Canaveral is 18.04 km
- The distance from CCAFS-LC-40 to the Atlantic Ocean is 0.93 km

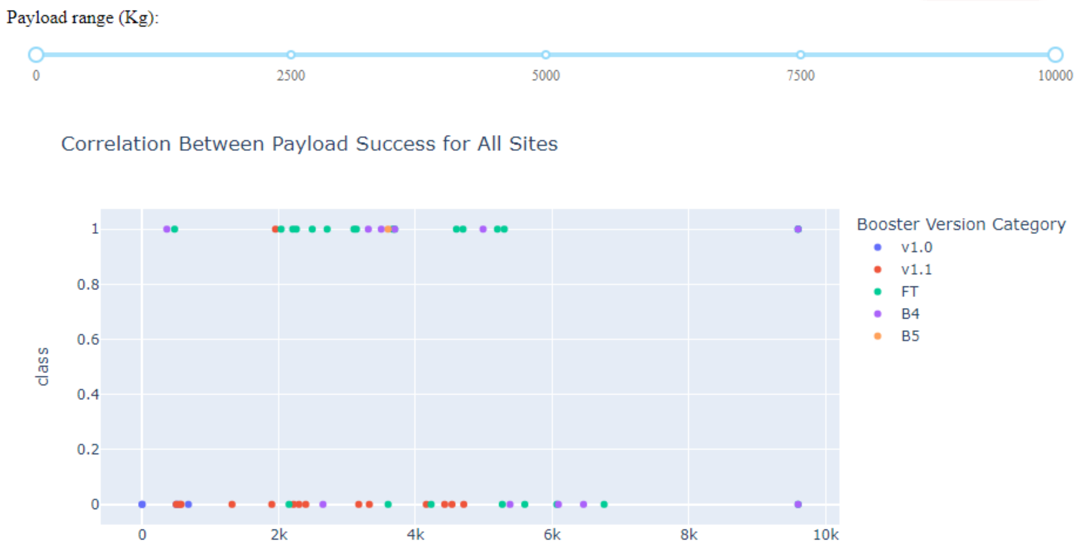
Interactive Dashboard

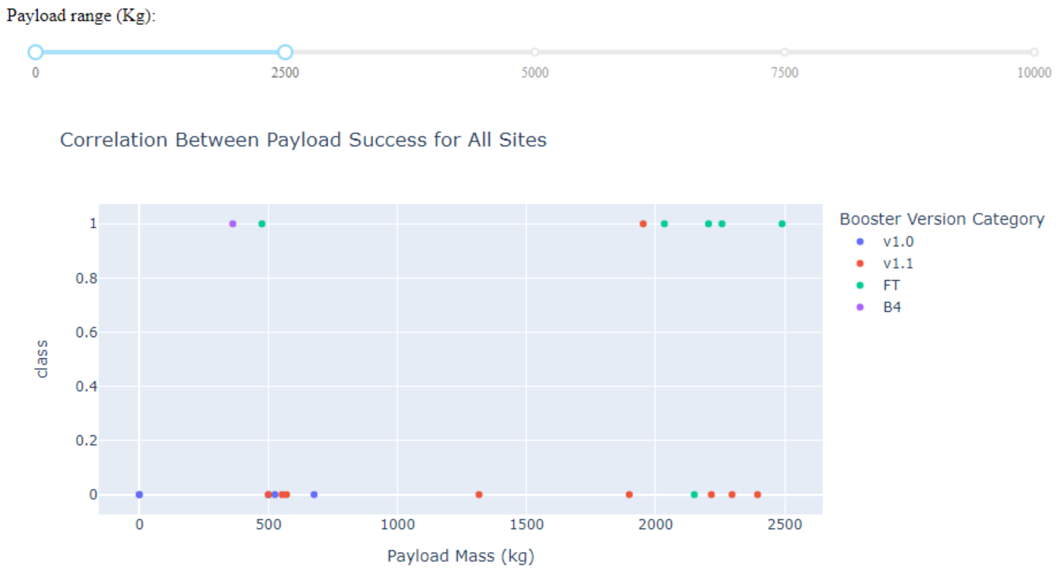




- Of the 4 launch sites, KSC LC-39A was used the most, 41.7% of the time
- CCAFS SLC-40 was used the least, 12.5% of the time
- KSC LC-39A had a 79% success rate on launches, which was the highest among the launch sites

Interactive Dashboard



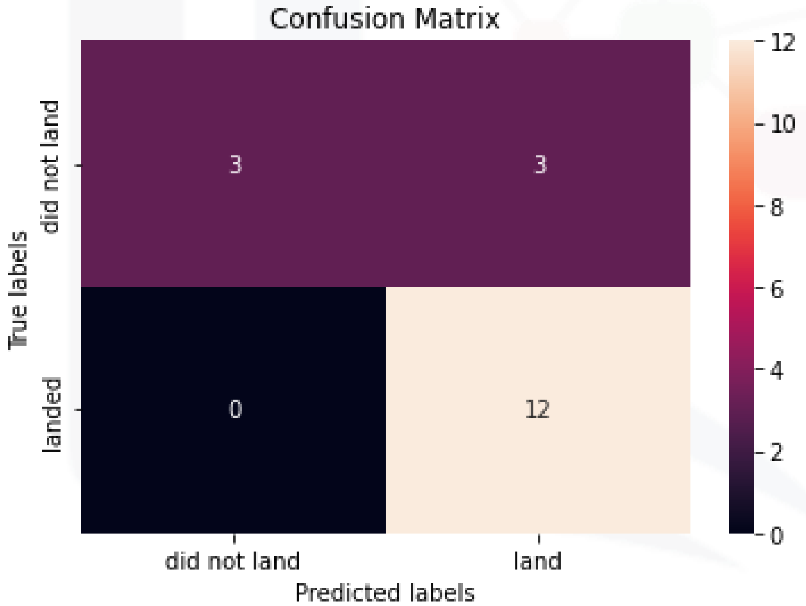


- For all payload weights, booster versions 1.0 and 1.1 were not very successful
- The largest payloads have only been carried by boosters v1.0 and FT

- For payloads below 2500 kg, v1.0 had no successes, while BT v1.1 only had a single success each
- Booster version FT has only one failure with a payload below 2500 kg

RESULTS: Predictive Analysis

ML Algorithm	R-squared score
Logistic Regression	0.8333
SVC	0.8333
Decision Tree	0.8333
KNN	0.8333



- LR, SVC, and Decision Tree and KNN all performed identically
- This identical performance is likely due to such a small testing set

- The confusion matrix looked identical for all of the algorithms
- The algorithms each predicted 3 false negatives, while they never correctly predicted a false positive

CONCLUSION

To generalize what was shown in this report, visual analysis showed us that success rate generally increased over time.

We also saw how the interactive dashboard explained that launch site KSC LC-39A was the most used, and most successful

Folium was used to explore the geospatial data relating to the launch sites, and SQL was used for preliminary EDA.

Lastly, we fit the training data to four different machine learning algorithms and saw identical performance from all of them.

APPENDIX

Figure A: Orbit Type vs. Payload Mass Scatter Plot

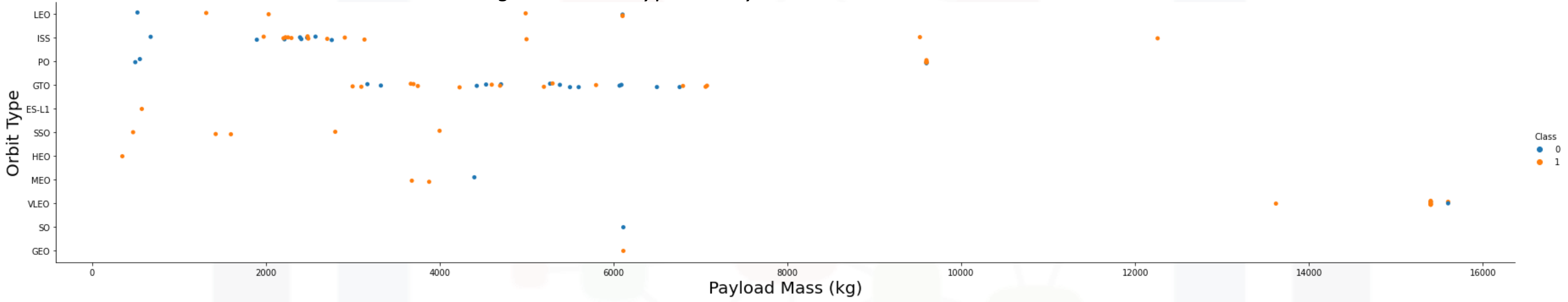


Figure B: Orbit Type vs. Flight Number Scatter Polt