Using Data Science to increase the probability that commercial space flights land successfully.

Name: Michael Lozos

Date: 9/3/2024

Outline

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

Executive summary/abstract

- Data from rocket launches were collected from application programming interfaces (API’s)
- API website is https://api.spacexdata.com/v4/launches/past
- Exploratory Data Analysis (EDA) with SQL
- Data Analysis Visualization with Pandas and Matplotlib
- Plotly Dashboards that allow users to interact with and modify real time data
- Folium for marking geographic locations and proximities of launch sites
- Machine learning algorithms such as logistic regression for predictive analysis

Executive summary/abstract continued

- Later flights, flights with larger numbers were more successful than earlier flights
- ES-L1, SSO,HEO and GEO orbit types all had perfect success rates, while SO orbit type had no successful launches.
- Flights with higher payloads, especially above 10,000kg were more successful than launches with lower payloads
- CCAFS LC-40 launch site had highest success rate
- Logistic Regression model was best at predicting whether the launch would succeed or not
- Booster type FT had highest success rate due to higher payload
- The only reason later flights, orbit types ES-L1, SSO,HEO and GEO, CCAFS LC-40 launch site had higher success rate was because the FT booster type was being used.

Methodology

Methodology

Executive summary

- Data Collection
  - Data is collected by API’s through the website https://api.spacexdata.com/v4/launches/past
  - A get request is used to retrieve data from the website. Data is stored as json objects
  - A Beautiful Soup object is used for web scrapping specific data from the website

Methodology

Executive summary continued

- Data Wrangling/Cleaning
  - Like with Data Collection, API's are used to wrangle data. Data is stored in Python lists
  - Data sample for Falcon 9 only was performed by extracting the Wikipedia website for Falcon 9 https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
  - Falcon 9 data was then cleaned using get requests and an Beautiful Soup object
  - Null values were eliminated by finding the mean of all Payload mass values. The Null is then replaced by the mean value.

Github URL= https://github.com/mwlozos24/Michael-Lozos

Methodology

Executive Summary continued

- Performed exploratory data analysis (EDA) using SQL, pandas, and matplotib
- Used interactive visual analytics using Plotly Dash for real time data interaction and Folium for marking precise geographic locations of launches and proximities of launch sites.
- Used machine learning algorithms such as logistic regression for predict analysis with test train splits.

Data Collection

- The Falcon 9 launch data was extracted by performing a get request on a the website https://api.spacexdata.com/v4/launches/past
- The get request only extracts raw unclean data

Data Collection API Call

Jupyter code that collects raw data from

Falcon 9 Rocket Launches

Github URL = https://github.com/mwlozos24/Michael-Lozos

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

Check the content of the response

```
print(response.content)
```

Data Collection Scraping

- Jupyter code for web scraping Falcon 9 launch with Beautiful Soup Object

Github URL= https://github.com/mwlozos24/Michael-Lozos

[35]:
```
# use requests.get() method with the provided static_ur
# assign the response to a object
Falcon_9 = requests.get(static_url).text
```

Create a `BeautifulSoup` object from the HTML `resp`

[38]:
```
# Use BeautifulSoup() to create a BeautifulSoup object
soup=BeautifulSoup(Falcon_9, 'html.parser')
```

Print the page title to verify if the `BeautifulSoup` obj

[39]:
```
print(soup.title)
```

```
<title>List of Falcon 9 and Falcon Heavy launches - Wik
```

Data Wrangling

Raw data from Falcon 9 launches was converted to pandas dataframe

Github URL= https://github.com/mwlozos24/Michael-Lozos

Load Space X dataset, from last section.

```
df=pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/dataset_pa
df.head(10)
```

Exploratory Data Analysis (EDA) with Data Visualization

- Scatter plot of flight number vs payload mass of rockets and whether the launches succeeded or failed
- Scatter plot of flight number vs launch site of rockets and whether the launches succeeded or failed
- Scatter plot of Launch Site vs payload mass of rockets and whether the launches succeeded or failed
- Bar graph of orbit type and Success rate to see what orbit type had highest success rate
- Scatter plot of orbit type vs flight number and whether the launches succeeded or failed
- Scatter plot of orbit type vs payload mass of rockets and whether the launches succeeded or failed
- Line plot of success rate vs time to see if success rates increased or decreased over time

EDA with SQL Queries

- Performed SQL Query to find every launch site
- Performed SQL Query that displayed 5 launches at either CCAFS LC-40 or CCAFS SLC-40 sites
- Performed SQL Query that computed the total payload mass of all rockets
- Performed SQL Query that computed the average payload mass of all rockets
- Performed SQL Query that retrieved the date of the first successful ground pad landing
- Performed SQL Query that named all of the boosters that had successful launches and had a payload mass between 4000 kg and 6000 kg
- Performed SQL Query that listed the number of all successful and unsuccessful outcomes
- Performed SQL Query that listed all of the booster names that had the maximum payload mass.
- Performed SQL Query that listed all failed landings in the year 2015
- Ranked all of the landing outcomes in descending order between the dates 6-4-2010 and 4/21/2017

Map of launch sites with Folium

- Used Circles, Markers, Markerclusters, and Polylines as objects
- Used Circles to pinpoint precise geographical coordinates of the 4 rocket launch sites
- Used Markers so that users can visually see the location of the 4 rocket launch sites
- Used Markercluster objects to distinguish from successful and failed launches at the same site
- Used Polylines to measure distance between launch sites and coastlines

Dashboard with Plotly Dash

- Pie Chart to determine which launch site had the highest percentage of successful launches
- 4 additional Pie Charts for percentage success rate for each launch site
- Scatter plots of Mass of rocket payload vs the probability of success rate for each of the 4 launch sites
- Scatter plots determine whether the payload mass or launch site was a better predictor of a successful launch

Predictive Analysis

- Used Logistic Regression model, Support Vector Machine model, Decision Tree model, and k neighbors model to determine which model is most accurate.
- Used GridSearch object for all 4 models
- Used fit method to train all data in data sets
- Used score method to determine accuracy of all Test data and Machine Learning models
- Used confusion matrix to determine if false negatives or false positive are the bigger problems

Predictive Analysis Code

Code to produce Machine Learning Algorithms

Github URL= https://github.com/mwlozos24/Michael-Lozos

TASK 5

Calculate the accuracy on the test data using the method

```
logreg_cv.score(X_test,Y_test)
```

```
0.8333333333333334
```

Lets look at the confusion matrix:

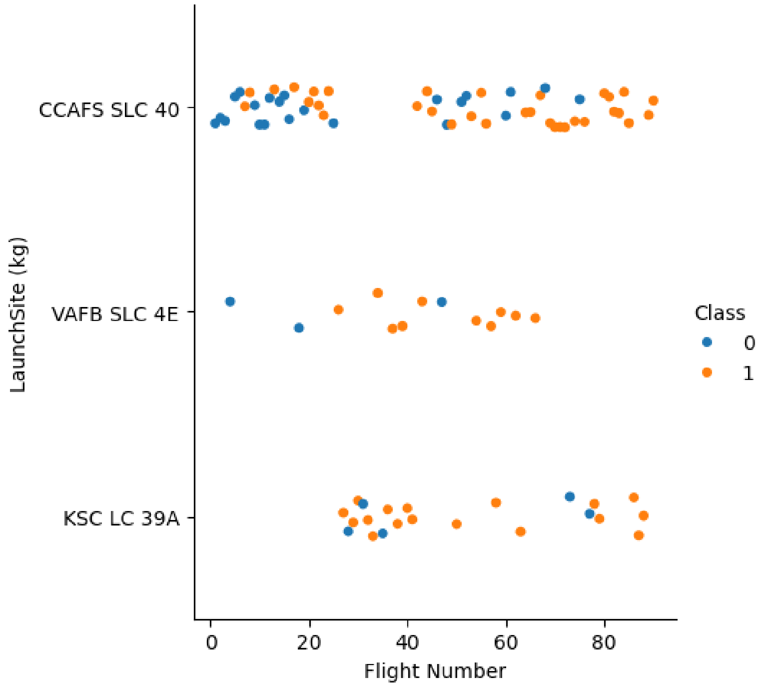
```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

Results

- Results for SQL queries, visual data analysis, Plotly Dashboard, Folium maps, and machine learning models/algorithms

Results and insights for exploratory data analysis

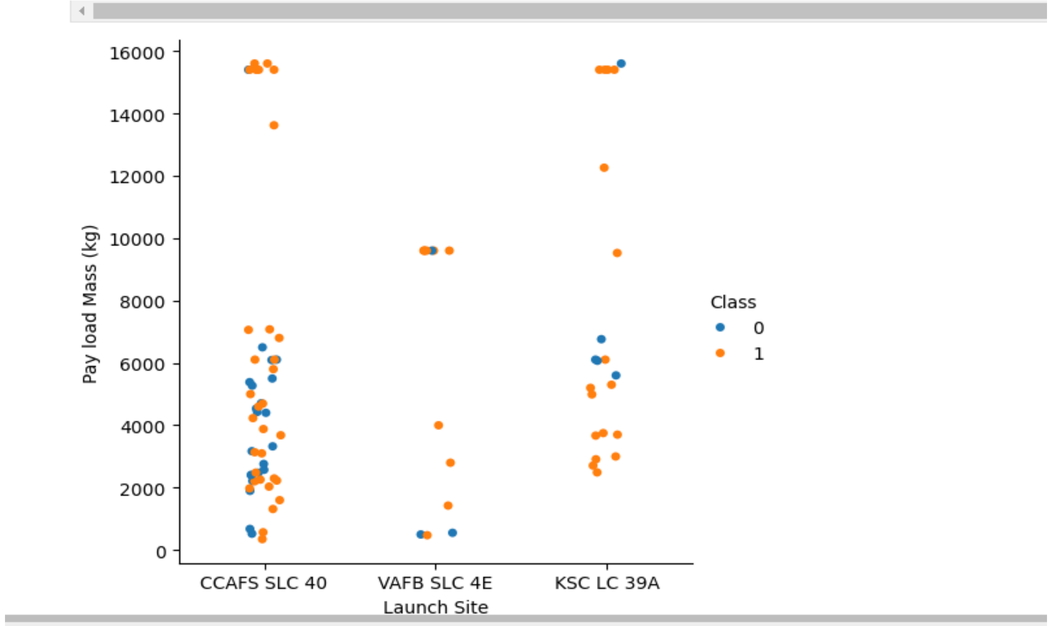
Flight number vs Launch Site



- Scatter plot indicates later flights, ones with higher numbers, had a higher probability of successful landings. This was especially true at the CCAFS SLC 40 site.

Github URL=
https://github.com/mwlozos24/Michael-Lozos

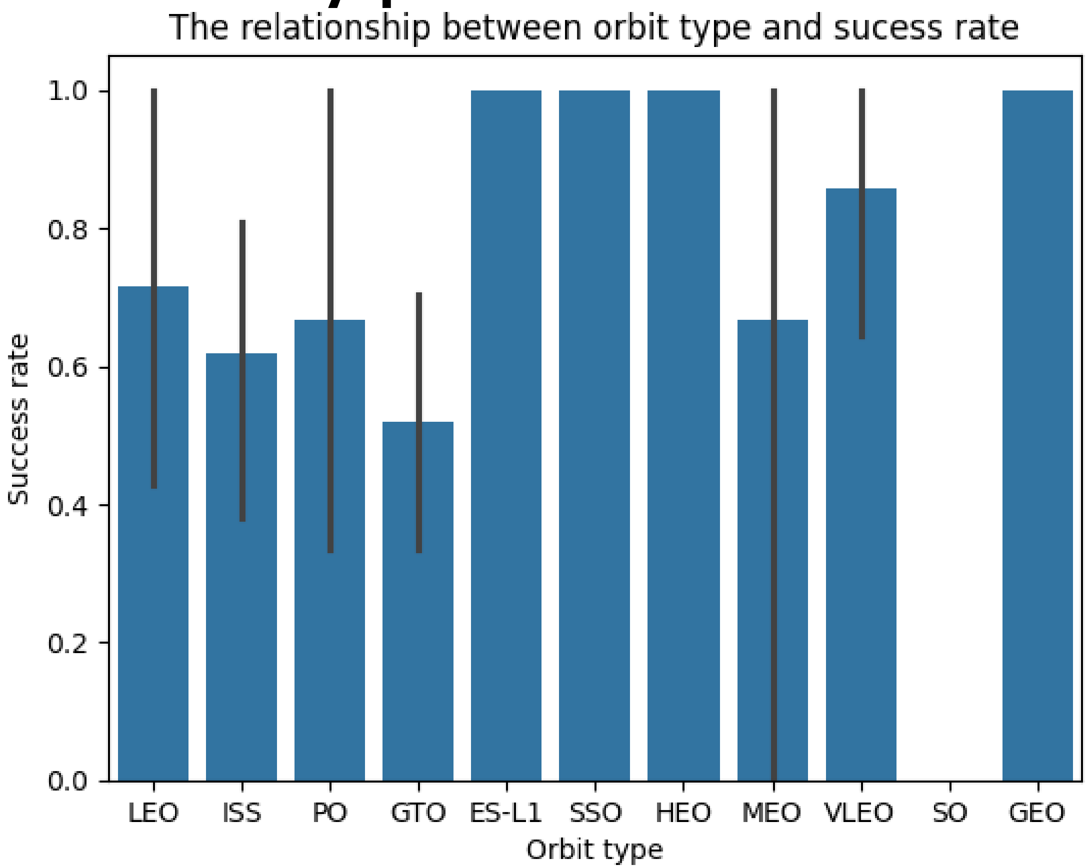
Payload mass vs launch site



- Scatter plot indicates probability of successful landing drastically increase if payload mass is greater 10000 kg. This is especially true for the CCAFS SLC 40 site. At the VAFB SLC 4E site there were no payloads greater 10000 kg.

Github URL=
https://github.com/mwlozos24/Michael-Lozos

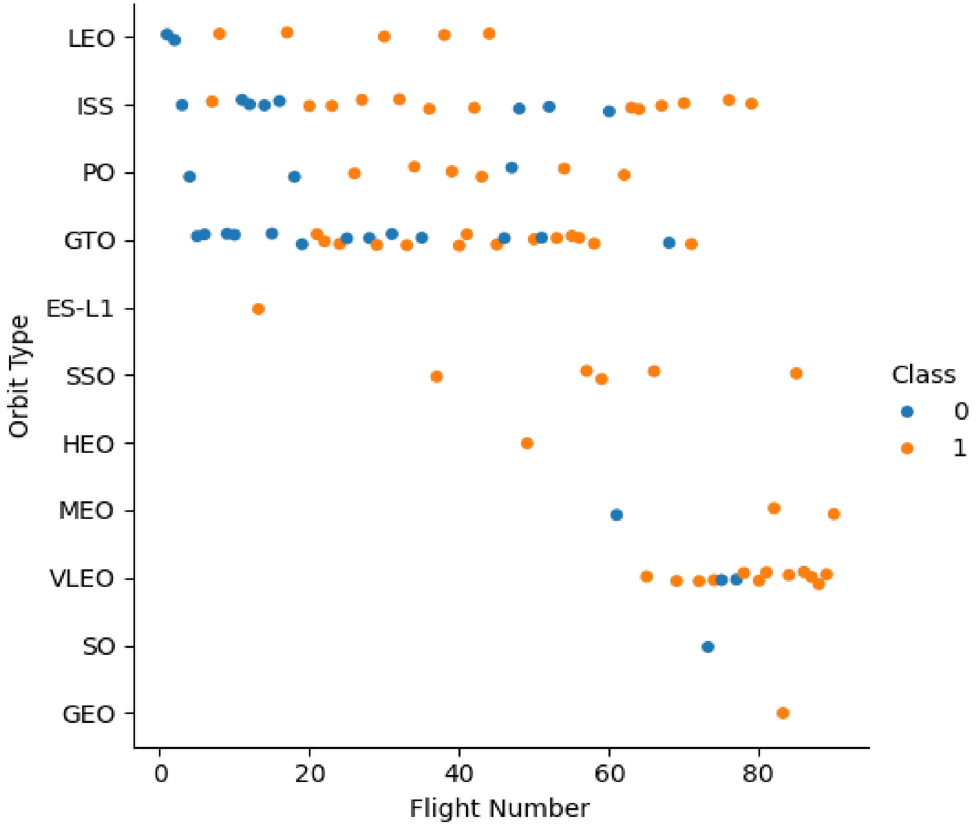
Orbit type and Success rate



- Orbit types ES-L1, SSO,HEO and GEO all had perfect 1.0 success rates.
- Orbit type SO had no successful launches.

Github URL= https://github.com/mwlozos24/Michael-Lozos

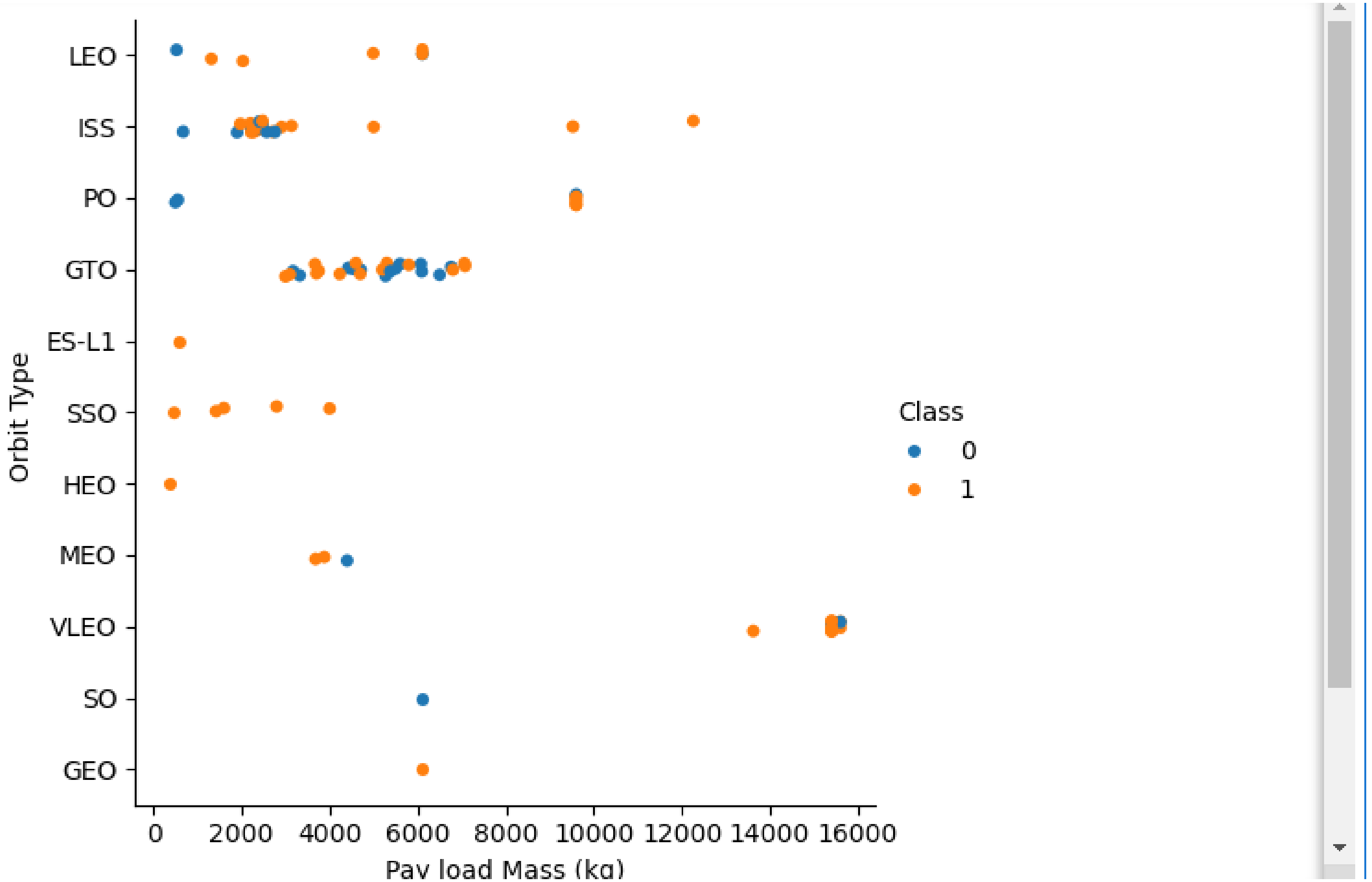
Orbit type and flight number



- It does not matter the orbit type later flights, ones with higher numbers, had more successful landings
- The exception is the SO orbit type, which one had launch and one failed landing
- Later flights with VLEO orbit type look especially promising

Github URL= https://github.com/mwlozos24/Michael-Lozos

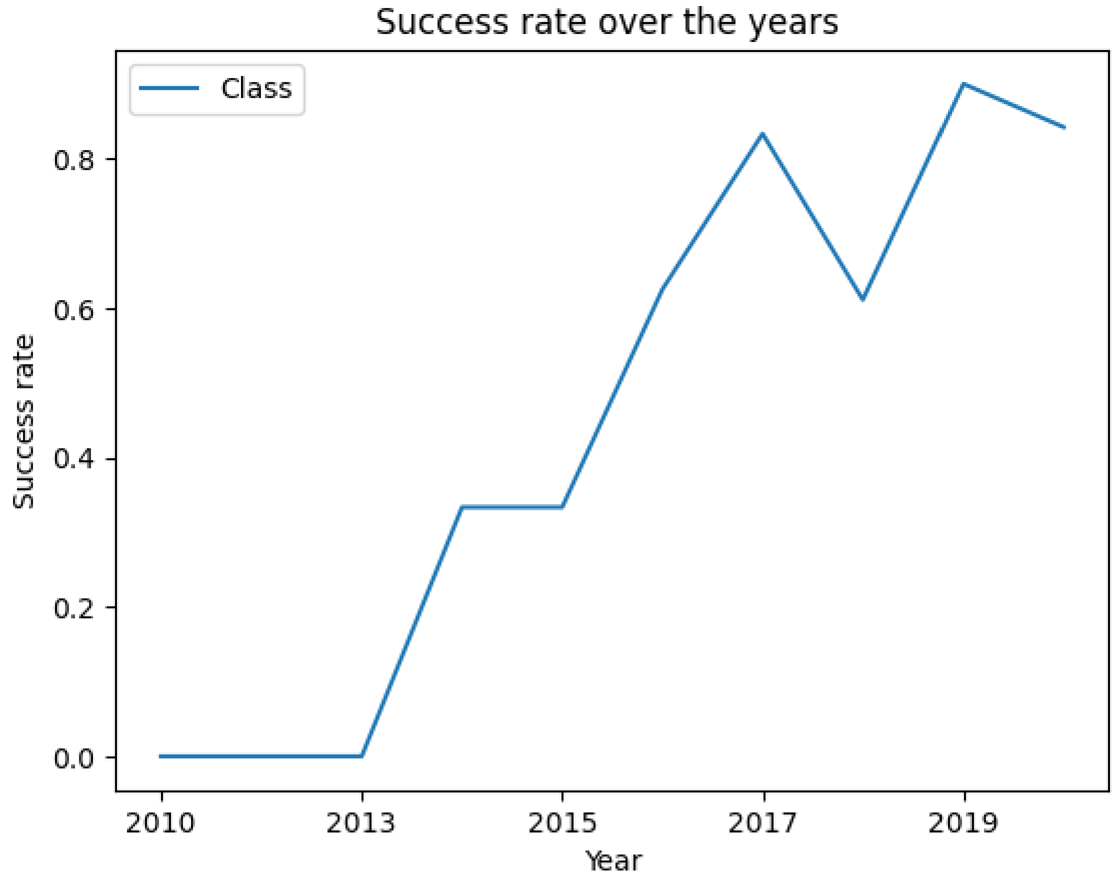
Payload Mass and Orbit type



- Perhaps the reason the VLEO orbit type had such a high success rate was payloads were greater than 10000 kg
- ISS had one launch above 10000kg and it too had a successful landing

Github URL= https://github.com/mwlozos24/Michael-Lozos

Success rate over the years



- Line graph clearly indicates success rate of rocket launches increased over time

Github URL= https://github.com/mwlozos24/Michael-Lozos

SQL query of all launch sites

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

- There are a total of 4 launch sites CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40
- CCAFS LC-40 is in Cape Canaveral, FL, , VAFB SLC-4E is in Vandenberg, CA, KSC LC-39A is Kennedy Space Center, FL CCAFS SLC-40 is a newer space station in Cape Canaveral Florida

Github URL= https://github.com/mwlozos24/Michael-Lozos

Querying 5 names that begin with CCA

Date	Time (UTC)	Booster_Version	Launch_Site
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40

- Both CCAFS LC-40 and CCAFS SLC-40 start with CCA but this query is in chronological order. The first 5 rocket launches that start with CCA were all at CCAFS LC-40.

Github URL= https://github.com/mwlozos24/Michael-Lozos

Total payload mass carried by NASA boosters

sum(PAYLOAD_MASS__KG_)
45596

- Total payload mass carried by NASA (CRS) boosters only

Github URL= https://github.com/mwlozos24/Michael-Lozos

Average payload mass for F9 v1.1 booster

avg(PAYLOAD_MASS__KG_)
2928.4

- Average payload mass for F9 v 1.1 boosters only

Github URL= https://github.com/mwlozos24/Michael-Lozos

First successful ground pad landing

Mission_Outcome	First_Sucessful_Landing_Date
Success	2015-12-22

- First successful ground pad landing was on December 12^{th}, 2015

Github URL= https://github.com/mwlozos24/Michael-Lozos

All boosters with successful drone ship landings with payload mass between 4000 and 6000 kg

Booster_Version	PAYLOAD_MASS__KG_	Landing_Outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

- Booster versions with successful drone ship landings that had a payload mass between 4000 and 6000 Kg
- Include F9 FT B1022, F9 FT B1026, F9 FT B1026, F9 FT B1021.2, and F9 FT B1031.2

Github URL= https://github.com/mwlozos24/Michael-Lozos

Mission outcomes

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

- Out of 100 rocket launches there were 99 successful missions and 1 failed mission
- One successful rocket launch had unclear payload status

Github URL= https://github.com/mwlozos24/Michael-Lozos

Booster's with maximum payload mass

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 boosters with a 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 B1051.4, F9 B5 B1049.5, F9 B5 B1060.2, F9 B5 B1058.3, F9 B5 B1051.6,
- F9 B5 B1060.3, and F9 B5 B1049.7,
- Github URL= https://github.com/mwlozos24/Michael-Lozos

Failed drone ship landings

Date	Landing_Outcome	Booster_Version	Launch_Site
2015-01-10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

- There were only 2 failed drone ship landings in 2015
- The first was on January 10th, 2015, The Booster Version was F9v1.1B1012, and the Launch site was CCAFS LC-40
- The second was on April 14th, 2015, The Booster Version was F9v1.1B1015, and the Launch site was CCAFS LC-40
  Github URL= https://github.com/mwlozos24/Michael-Lozos

All Landing Outcomes of success ground pad landing and failure drone between 6-4-2010 and 4/21/2017

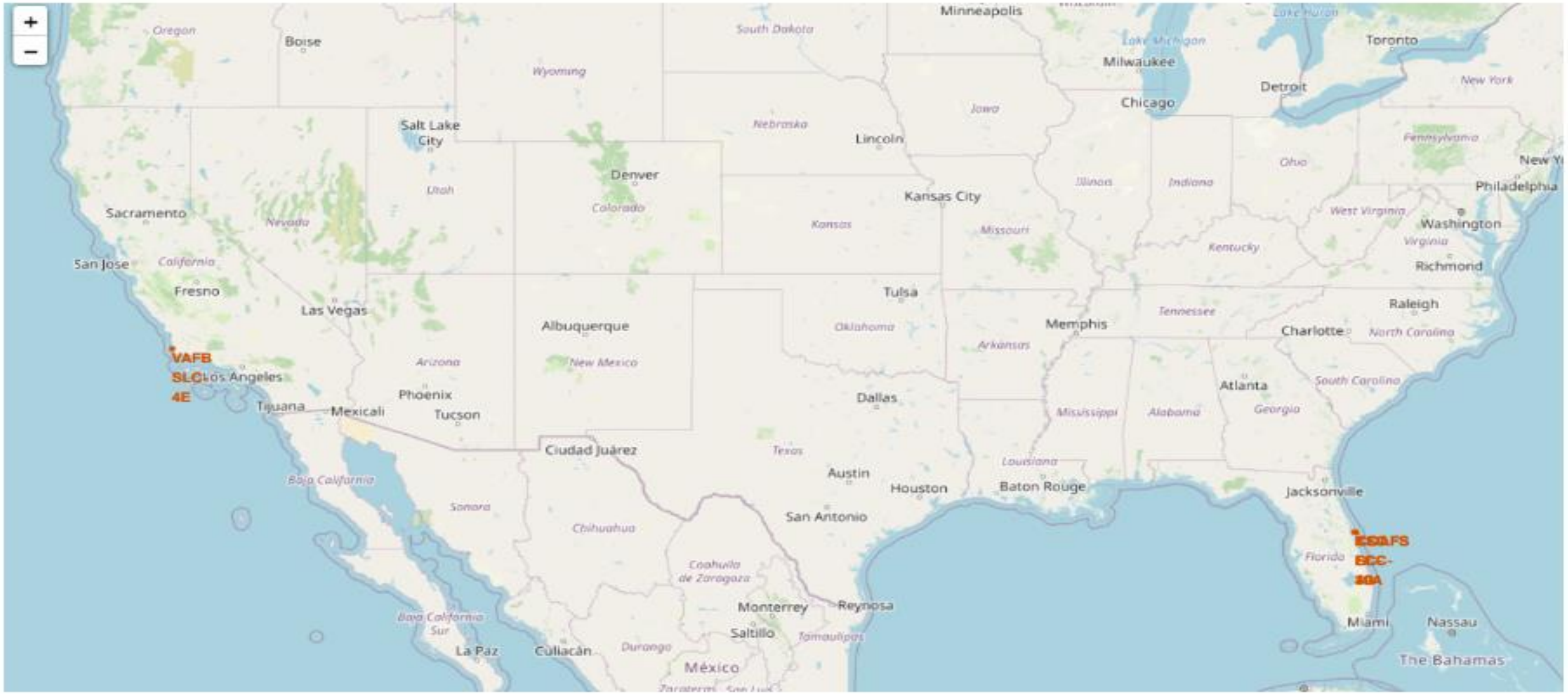
Date	Landing_Outcome	Total_Landings
2015-12-22	Success (ground pad)	3
2015-01-10	Failure (drone ship)	5

- There was a total of 3 successful ground pad landings on December 22nd, 2015
- There was a total of 5 failed drone ship landings on January 10th, 2010
- Query was done in reverse chronological order

Github URL= https://github.com/mwlozos24/Michael-Lozos

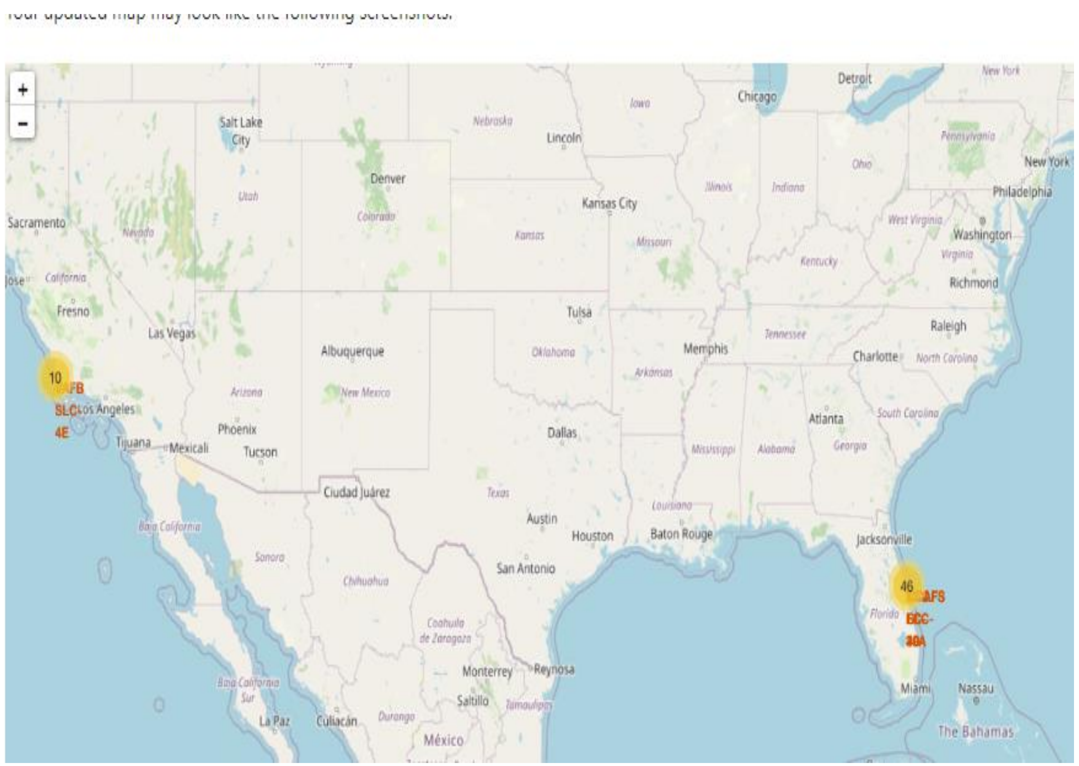
Launch Sites and proximities to coastlines highways, and cities

Folium map of all launch sites



- This map shows that all 3 launch sites are near the coast
- The CCAFS 40-LC and CCAFS 40-SLC sites are in east central Florida
- The VAFB SLC 4E is in southwest California northwest of Los Angeles
- All 3 sites are below 35 degrees latitude
- Github URL= https://github.com/mwlozos24/Michael-Lozos

Folium map on launch site outcomes



- Red indicates failed launches
- Green indicates successful launches
- 26 indicates the total number of failed launches

- This maps shows there was a total of 56 rocket launches
- 10 were at the KSC LC-39A in California
- 46 were at the CCAFS SLC-40 and CCAFS LC-40 sites in Florida

Distance to Coast, Highway, and City

Your updated map with distance line should look like the following screenshot:


```
Distance to closet highway is 0.1654425465951123 km
Distance to closet railway is 1.5798401528109256 km
Distance to closet city is 52.88719788047493 km
```

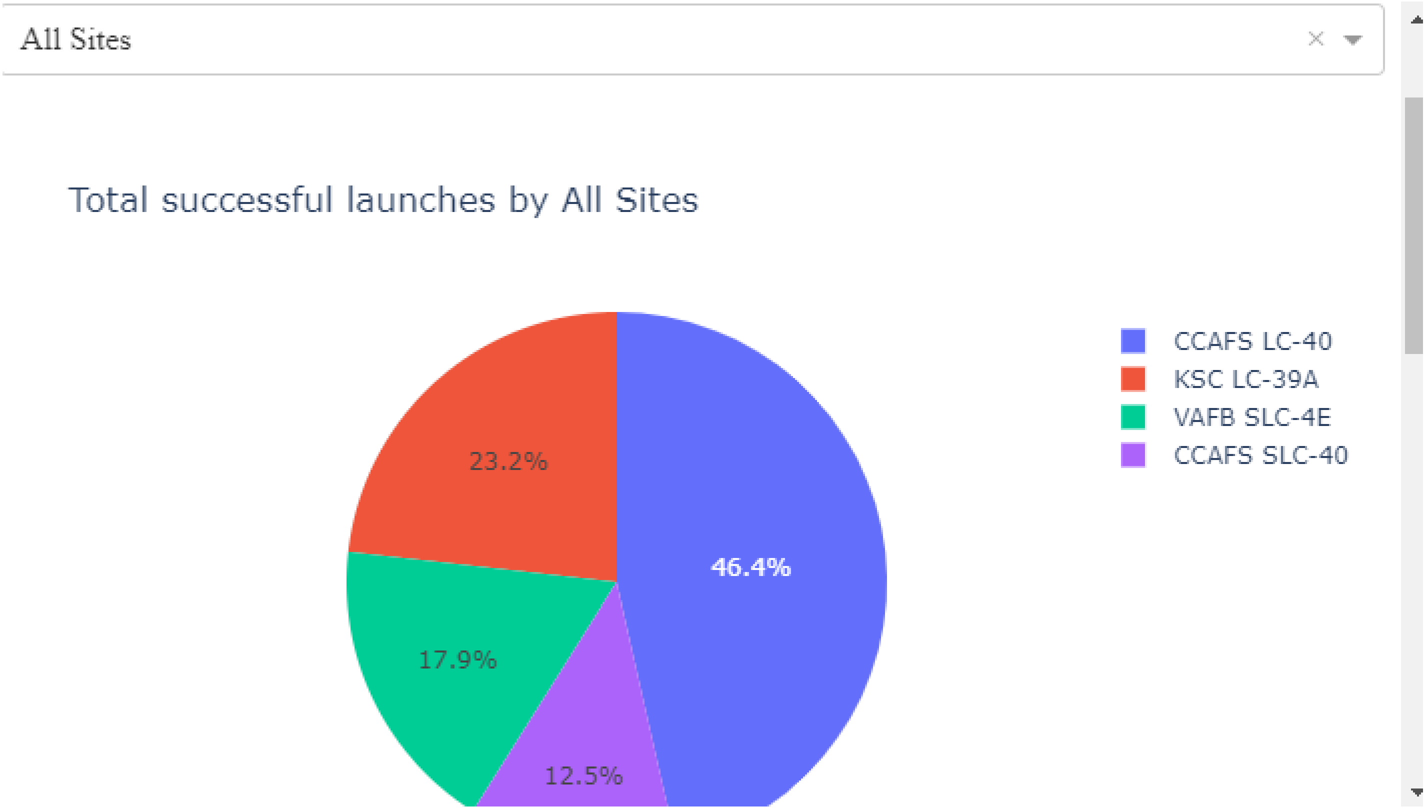
- Distance to closet highway was 1.66 kilometers
- Distance to closet railway is 1.58 kilometers
- Distance to closet city is 52.89 kilometers

- Distance to coast from CCAFS SLC-40 site was only .91 kilometers

Github URL= https://github.com/mwlozos24/Michael-Lozos

Dashboard

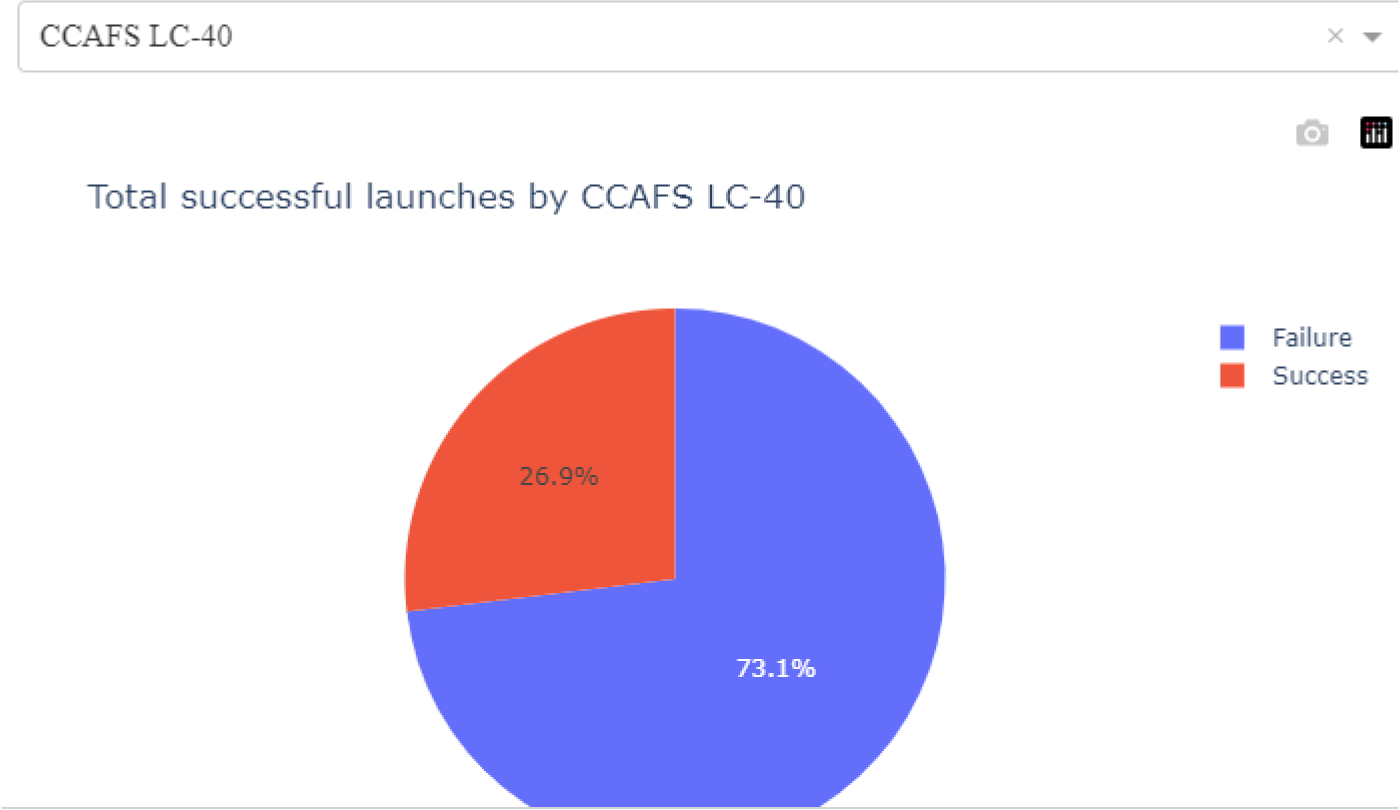
Plotly dashboard of success rates by launch site



- Pie chart indicates CCAFS LC-40 site had highest success rate at 46.4%
- KSC LC-39A was next highest at 23.2%
- VAFB SLC-4E was third highest at 17.9%
- CCAFS SLC-40 had lowest success rate at 12.5%

Github URL= https://github.com/mwlozos24/Michael-Lozos

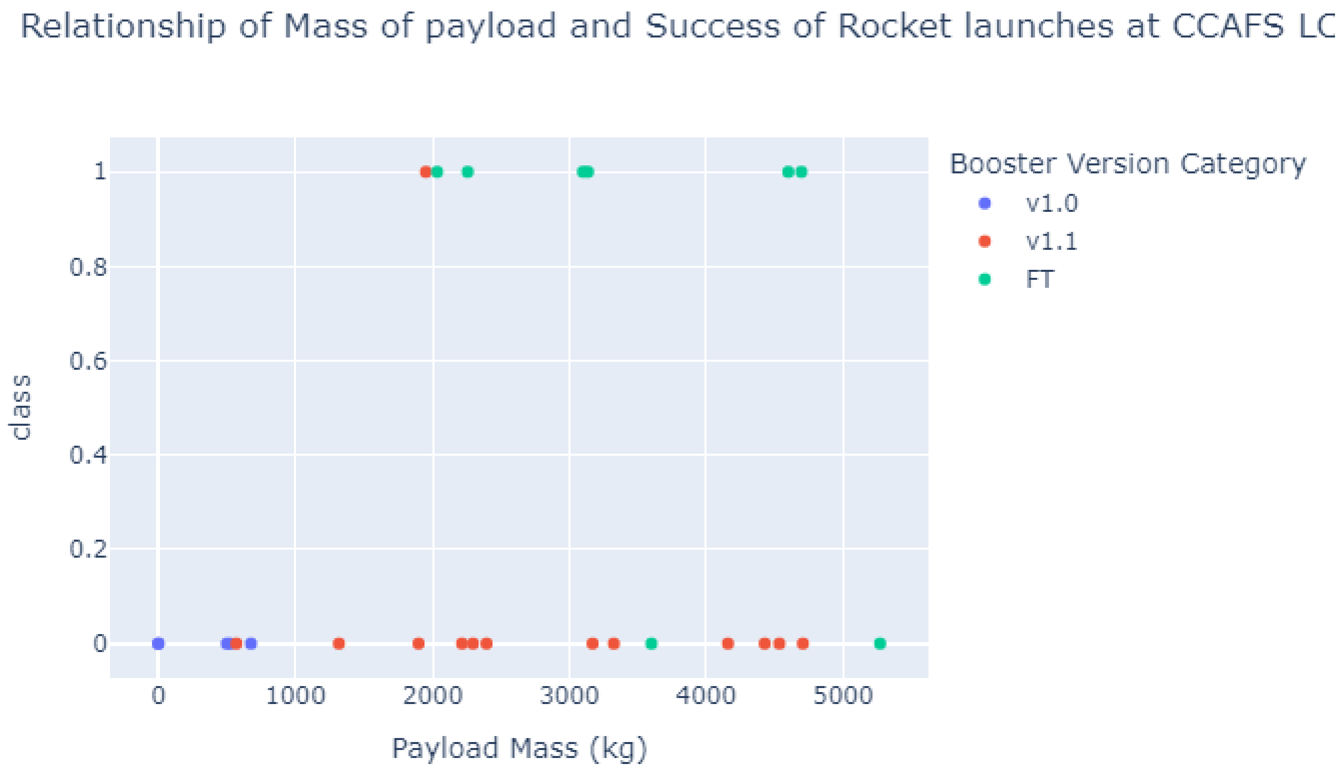
Success rate of CCAFAS LC-40



- This pie chart indicates that the CCAFS had a Success rate of only 26.9%
- This was much lower than 46.4% indicated by the pie chart with the All sites input

Github URL= https://github.com/mwlozos24/Michael-Lozos

The effect of booster type and payload on success rate

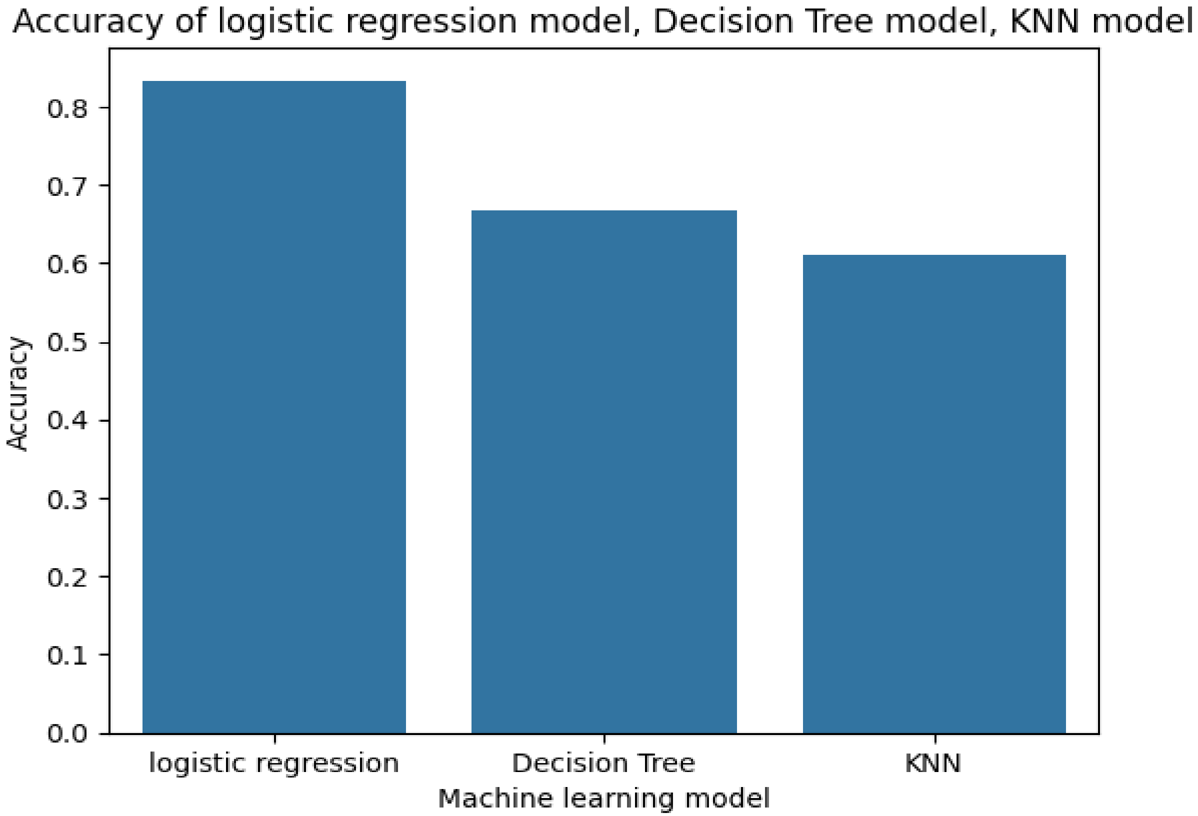


- Rockets with payloads below 2000 kg had no successful launches
- Booster version v1.0 had no successful launches
- Once payload was over 2000kg there was no correlation between increasing payload mass and success rate
- v1.1 booster was successful on only one launch attempt while FT booster was successful 75% of the time

Github URL= https://github.com/mwlozos24/Michael-Lozos

Machine Learning Predictive Analysis

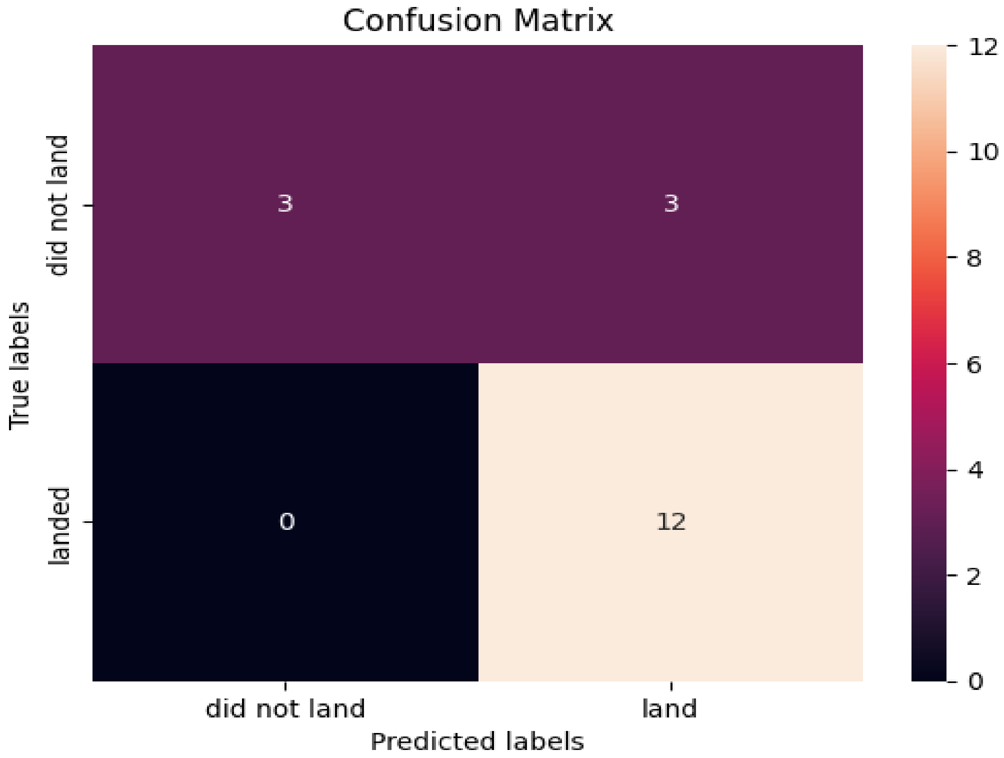
Highest performing machine learning model



- Logistic regression had the highest accuracy of machine learning models at 83.33%
- Decision Tree Machine learning model was next highest performing at 66.67%
- K Nearest Neighbors was worst performing machine learning algorithm at only 61.11%
- Support Vector Machine (SVM) was omitted from dataset because data from training model didn't load in time

Github URL= https://github.com/mwlozos24/Michael-Lozos

Confusion Matrix for Logistic Regression



- Confusion indicates that false positives, which are rocket launches the Logistic Regression predicted to succeed actually failed was the problem with the Logistic Regression Model
- There were no false negatives. In other words if the Logistic Regression Model predicted a failed launch, then the launch actually failed
- Logistic Regression model was correct on 15 out of 18 predictions matching the 83.33% indicated by bar graph.

Github URL= https://github.com/mwlozos24/Michael-Lozos

Conclusions

- No matter the launch site later flights had a higher probability of succeeding than earlier flights
- Flights with higher payloads were more likely to succeed than lower payloads, especially payloads of over 10,000 kg
- Orbit types ES-L1, SSO,HEO and GEO all had 100% success rates.
- SO Orbit type had no successful launches
- CCAFS LC-40 launch site had the most successful launches
- FT booster type had the highest success rate
- Logistic Regression machine learning model was best at predicting whether rocket launches were successful or not. This is not surprising because the dependent variable is a binary succeed or fail and not a numerical data type.

Conclusion and further questions for future launches

- What variable was the predictor of a successful launch?
- Launch Site, Orbit Type, date, booster type, or some other variable?
- How can the Logistic Regression Model be improved to more accurately predict whether a launch will succeed or not?
- Booster type is most important variable
- FT booster has larger engine and fuel tank hence larger payload mass
- The reason for CCAFS SLC-40 low success rate is that it was using v1.0 and v1.1 boosters instead of FT boosters
- The reason success rates improved over time is FT boosters were used instead of v1.0 and v1.1 boosters

Conclusions continued

- Use Full thrust booster type started on December 22nd 2015 and success rate increased sharply afterwards not a coincidence.
- The reason for orbit types ES-L1, SSO,HEO and GEO perfect success is because it was after the full thrust booster type started being used and had nothing to do with orbit type
- Launch site only matters if it is close enough to coast and all 4 sites are very close to the coast. Only the date of launches matter because that determines whether FT boosters were used or not.
- Rocket launches are less likely to be successful with payload mass below 10000kg because rocket will lack engine power and fuel tank size. FT booster has maximum payload of 15600Kg
- Future Logistic Regression Machine Learning models should predict the success of rocket launches based on booster type and payload mass rather than launch site, orbit type, or date

Appendix

- This is proof that the Booster type was switched to
 FT or v1.2 on December 22nd, 2015.
Screenshot of NASA Datasheet for v1.2 or FT booster type.

- Notice the steep increase after 2015,
 this was because the booster type switched to FT or v1.2
On December 22nd, 2015

The End