Project Helios

Stage One Redeployment Feasibility

David Jolley

4-15-2022

Outline

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

Executive Summary

Background

- Star Aeronautics is a new company wanting to develop a rocket launch business where launches cost anywhere between $65M to $165M. Star Aeronautics asked us to see if we could develop a model the parameters of a successful Stage 1 rocket reuse based on current market test data that is available. This would be used to develop the new Helios rocket platform. This should drastically reduce the needed capital to start operations for the Helios program.

- In order to create a model to predict launch outcomes, data was gathered from an API and web scraping functions from publicly available data from SpaceX.

- The data was then cleaned and analyzed using SQL, Matplotlib and Seaborn visualizations. Plotly interactive dashboard and Folium geolocation Were also used to visualization certain parameters of the model. Logistic regression prediction models were created using Scikit-learn and the best models were chosen using confusion matrices and model accuracies.

Findings

- The SVM, Logistic Regression and KNN can predict the outcome of a launch for Star Aeronautics with 84% accuracy and 80% preci sion.

Recommendation

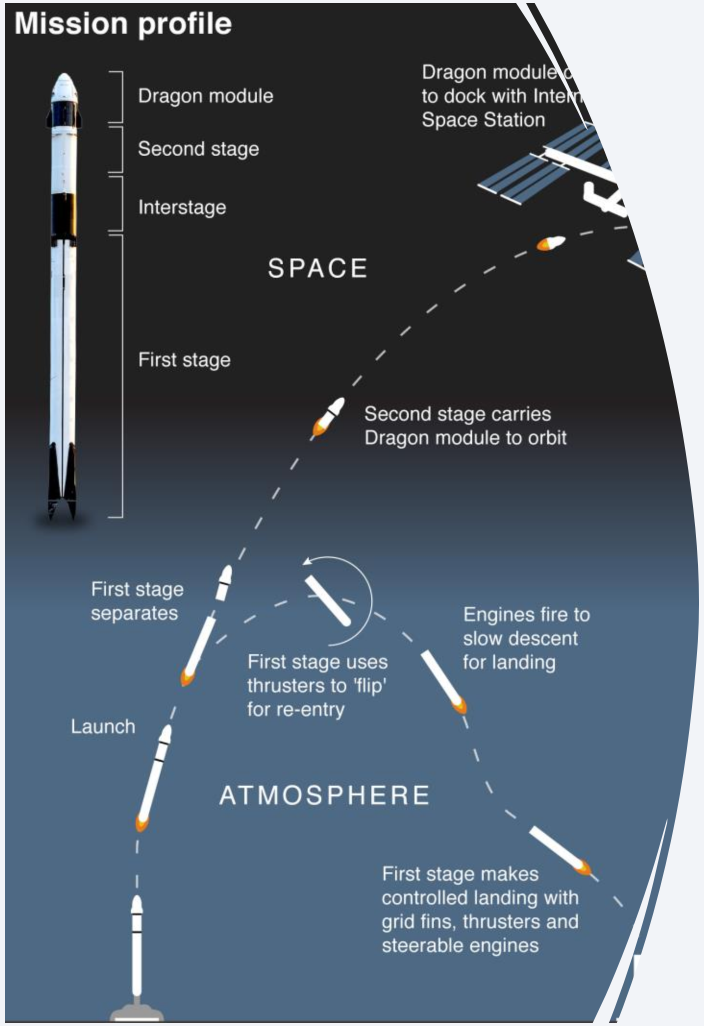
- There is a risk to price the launch into space with the provided prediction models because the model accuracy is at only 84%.

- Predicting the successful landing had much better results than predicting the failures.

- Star Aeronautics could lose $100M on each failed prediction. Since the cost of Phase 1 depends on the number of times it is reused (average cost), not whether the outcome of one launch is successful, Star Aeronautics should review the parameters selected before the Helios program it launched.

- A larger dataset could help improve accuracy and precision of these models.

Introduction



- Commercial rocket aeronautics has become a big market!
- Star Aeronautics is a new company and wants to compete with SpaceX on pricing.
- SpaceX advertises Falcon 9 rocket launches on its website for a cost of $62 million dollars. SpaceX costs are lower than other providers because they can recover the first stage of the rocket launch.
- Competitors to SpaceX advertise rocket launches for a cost of $165 million dollars.
- I will use machine learning to model whether SpaceX will reuse the first phase based on the parameters available. This will enable Star Aeronautics to price a launch for their entry of the Helios program into the current market.

Section 1

Methodology

Methodology

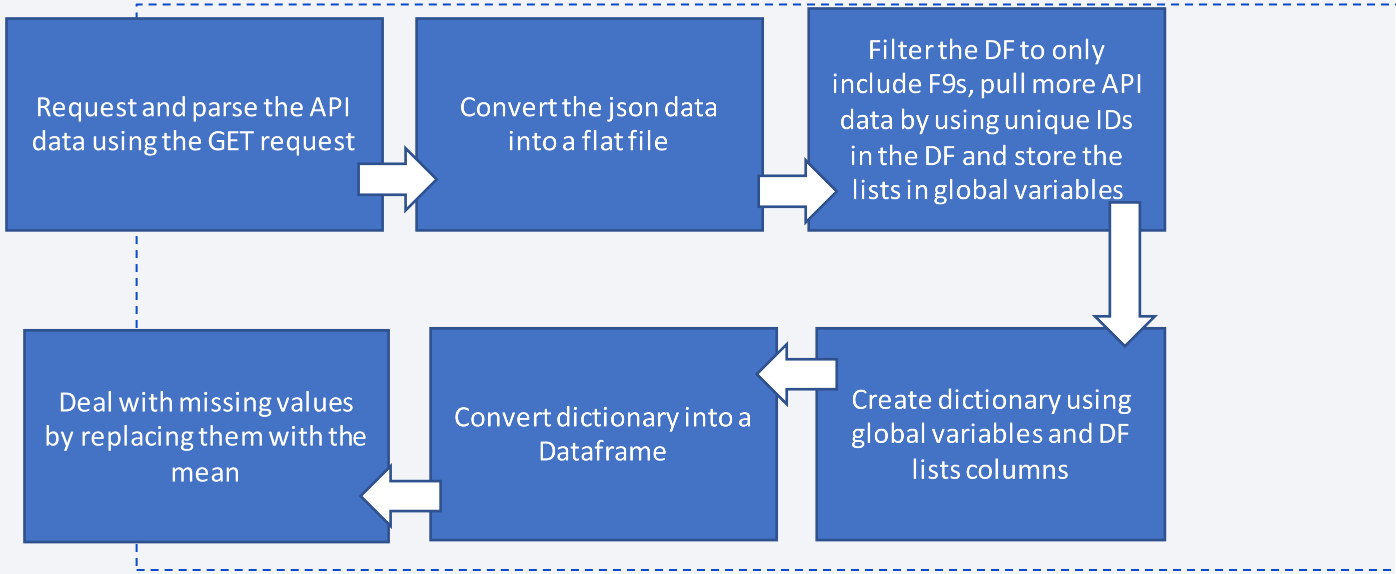
- Data collection methodology:
- Used get request with API and Beautiful Soup with Web Page to ingest data and create Pandas dataframes.
- Perform data wrangling
- Used Pandas and Numpy to transform the messy and incomplete data into actionable information for analysis.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
- Used Scikit-Learn to create LR, SVM, KNN and Decision Tree models and tune with GridSearchCV.

Data Collection

- API Gathered data on historical SpaceX launches using SpaceX API and the Get Command. The data in the API was filtered to only include Falcon 9 launches.
- Falcon 9s can lift heavier payloads.
- Web Scraping Gathered data on historical SpaceX launches using web scraping. The Web
- Scraping provided additional information not found in API including customer, time and booster version.
- To make the answers consistent with the API, the data was pulled in a pre-selected date range. (121 rows)
- High Level Data Collection Process/Goal
- Parsed each of the data sets, converted the data into a pandas dataframe and
- cleaned the data to be used in Exploratory Data Analysis

Data Collection – SpaceX API

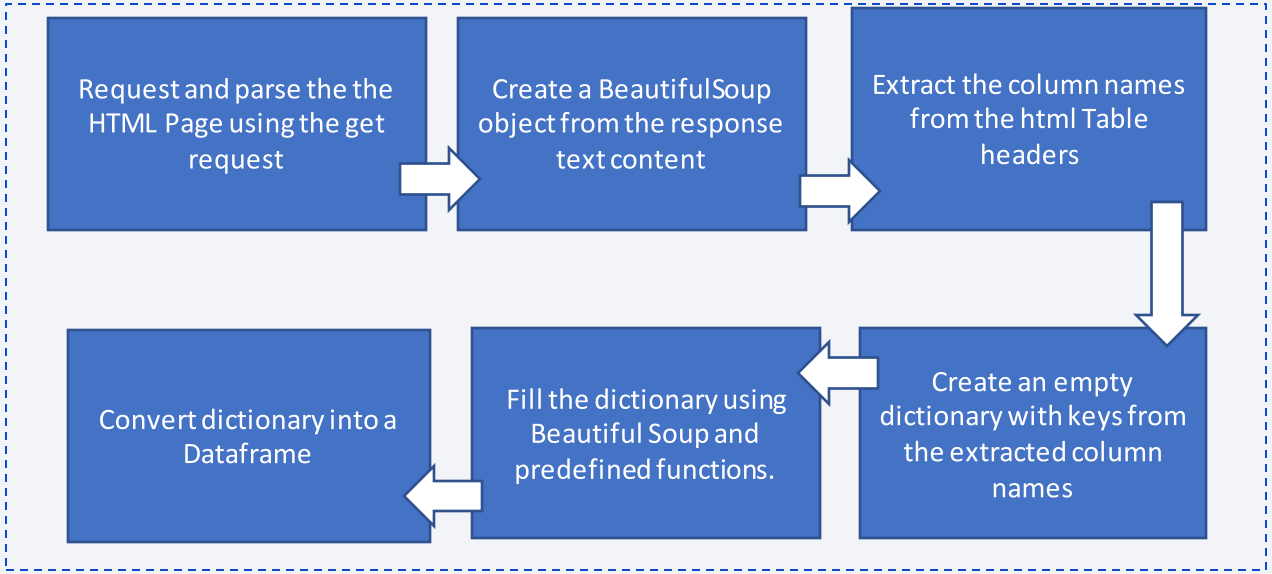
- Parsed API data, converted the data into a pandas dataframe and clean the data to be used in EDA
- GitHub URL for the Jupyter notebook: https://github.com/davidbjolley/Helios-Project/blob/main/Data%20Collection%20API.ipynb



	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedC
0	1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0
1	2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0
2	4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0
3	5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN	0
4	6	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0

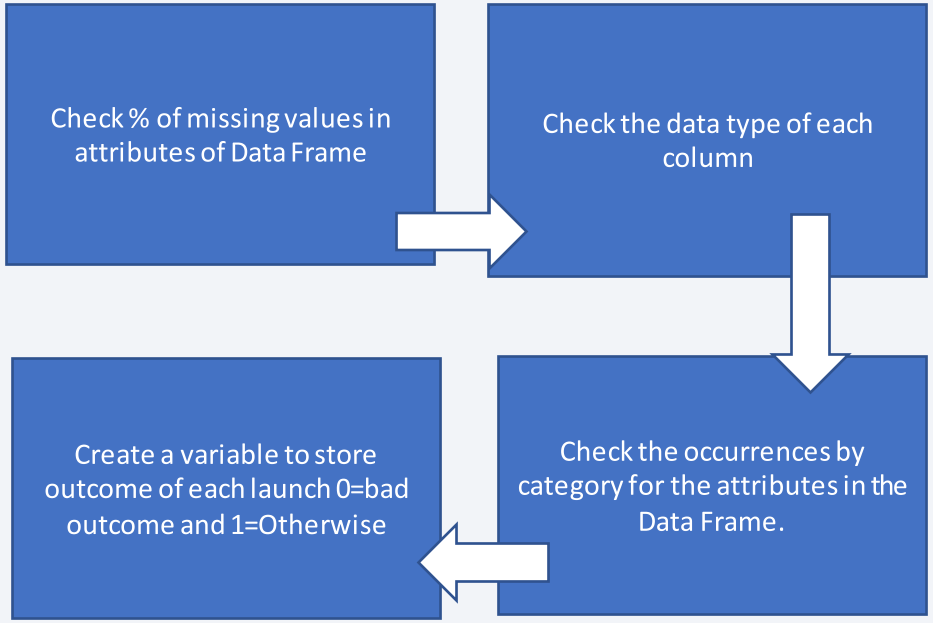
Data Collection - Scraping

- Parsed an HTML table from Wikipedia and converted it into a pandas dataframe to be used in EDA.
- GitHub URL for the Jupyter notebook: https://github.com/davidbjolley/Helios-Project/blob/main/Data%20Collection%20API.ipynb



Data Wrangling

- Using the API data, created the outcome variable (Y) with numeric value 0 or 1 that is necessary for creation of a classification models. The class labels are mapped to 1 for the positive class or outcome and 0 for the negative class or outcome. The models predict the probability that an example belongs to class 1.
- GitHub URL for the Jupyter notebook: https://github.com/davidbjolley/Helios-Project/blob/main/Data%20Wrangling.ipynb



EDA with Data Visualization

- Created data visualizations with the API data to examine the relationship between variables and outcome in order to select the features that will best predict in a machine learning mode
- Machine learning models require all input and output variables to be numeric. Therefore, features selected were encoded to numbers.
- GitHub URL for the Jupyter notebook: https://github.com/davidbjolley/Helios-Project/blob/main/EDA%20Visualization.ipynb

Scatterplot

1. Flight Number vs. Payload Mass
2. Payload vs. Launch Site
3. Flight Number vs. Orbit type
4. Payload Vs. Orbit type

Barplot

Success Rate (Mean) vs Orbit type

Line Plot

Launch Success yearly trend

EDA with SQL

- Imported the wikipedia web scraping data into DB2 and queried the data from Jupyter notebook to examine the relationship between variables and outcome in order to select the features that will best predict in a machine learning model

- GitHub URL for the Jupyter notebook: https://github.com/davidbjolley/Helios-Project/blob/main/EDA%20with%20SQL.ipynb

Distinct, Count and group
1. Names of the unique launch sites in the space mission
2. Landing outcome
3. Launch sites
4. Mission outcomes

Order, sum and average
1. Launch sites beginning with the string 'CCA'
2. Total payload mass carried by boosters launched by NASA(CRS)
3. Average payload mass carried by booster version F9 v1.1

Where clause and groups
1. The date when the first successful landing outcome in ground pad was achieved.
2. Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
3. Total number of successful and failed mission outcomes
4. Names of the booster_versions which carried the maximum payload mass.
5. Failed landing_outcomes in drone ship, their booster versions, and launch site names for the year 2015

Rank ○ Ranked 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

Built an Interactive Map with Folium

- Using the API data, mapped successful/failed launches by launch site and the distances of sites to proximities to be able to find some geographical patterns about launch sites in order to select the features that will best predict in a machine learning model

- GitHub URL for the Jupyter notebook: https://github.com/davidbjolley/Helios-Project/blob/main/Space%20Y%20Dashboard.ipynb

Objects added to the map

- Added folium.Circle and folium.Marker for each launch site on the site map
- Added a marker_cluster with green or red markers for every outcome by site and by launch to see which sites have high success rates
- Added a MousePosition on the map to get coordinate for a mouse over a point on the map so you use the coordinates to create a folium.Marker to show the distance
- Drew a PolyLine between a launch site and the selected coastline point

Built a Dashboard with Plotly Dash

- Using the Wikipedia information, added interactions to explore which site, payload range and booster version have the highest successful launches.

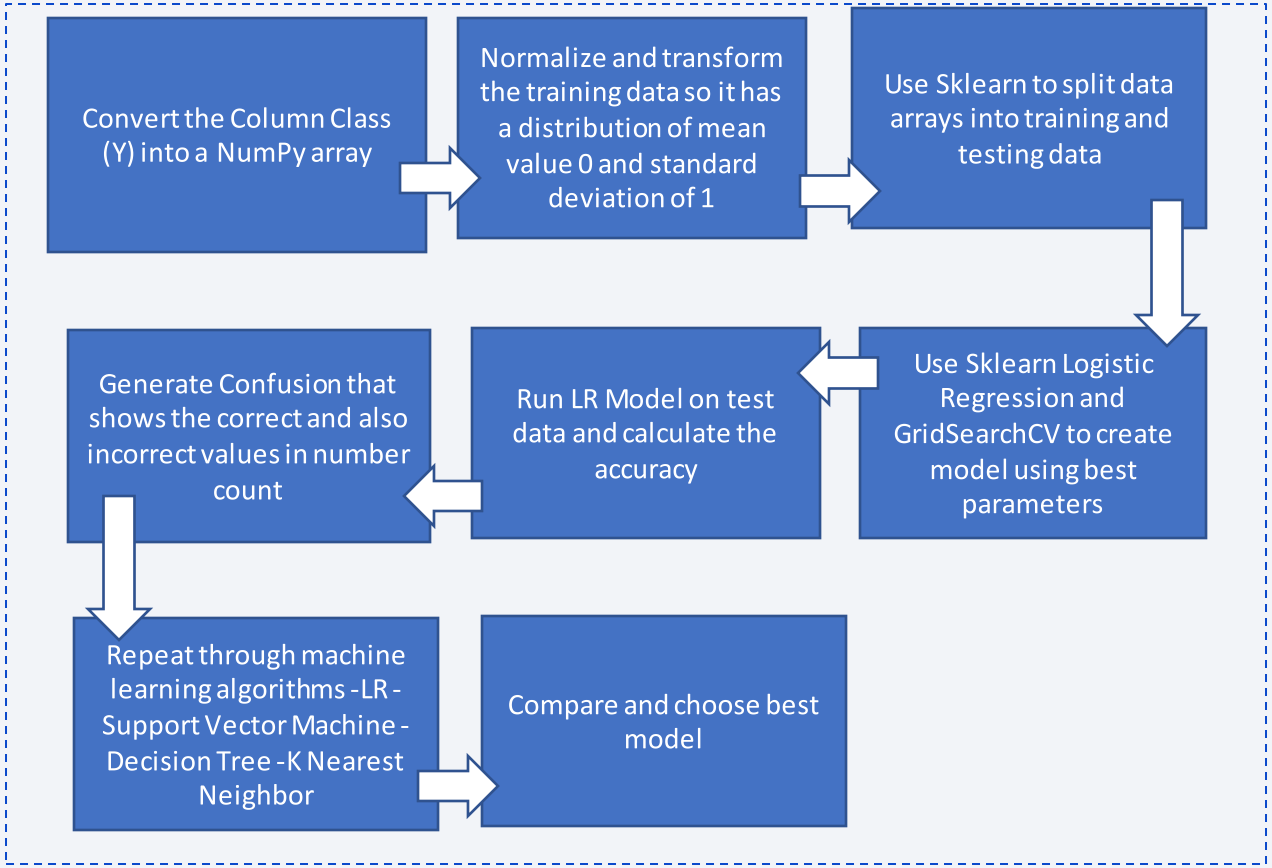
- GitHub URL for the Jupyter notebook: https://github.com/davidbjolley/Helios-Project/blob/main/spacex_dash_app.py

Plots/graphs and interactions added to a dashboard

- Dropdown menu so the user could choose a launch site or all sites.
- A pie chart driven by the drop-down menu that shows launch successes by site as a % of all launches and success rate for each individual site.
- An slider that enables choice of range of the amount of Payload mass he/she wants to analyze.
- A scatter plot driven by the slider that shows how the amount of Payload mass affects the results of the landing by booster version.

Predictive Analysis (Classification)

- Created machine learning classification models using Logistic Regression, SVM, Decision Tree and KNN
- Used GridSearchCV to choose the best Hyperparameters for SVM, Classification Trees and Logistic Regression
- Evaluated the models by calculating the accuracy and generating a confusion matrix.
- GitHub URL for the Jupyter notebook: https://github.com/davidbjolley/Helios-Project/blob/main/Machine%20Learning%20Prediction%20(1).ipynb



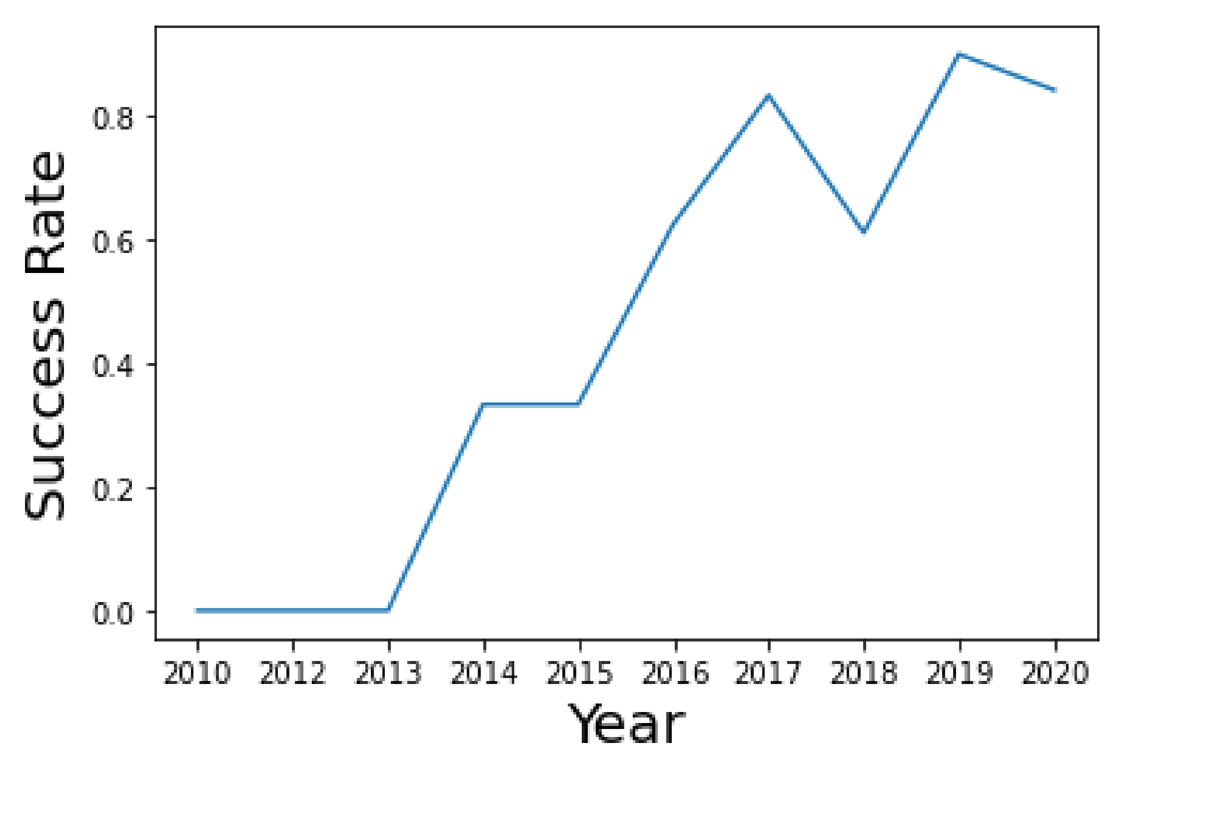
Results – Exploratory Data Analysis (EDA)

- Outcomes are improving over time represented by flight number.
  - Launches from CCAFS SLC 40 locations increasing over time
  - Launches from ISS and VLEO orbits increasing over time
- Payload Mass is increasing over time.
  - ISS, VLEO and Polar Orbits can successfully handle heavier payloads

- 12 Launches (out of 69) that were successful and had maximum payload of 15,600KG were by Booster Version version F9 B5 B10.

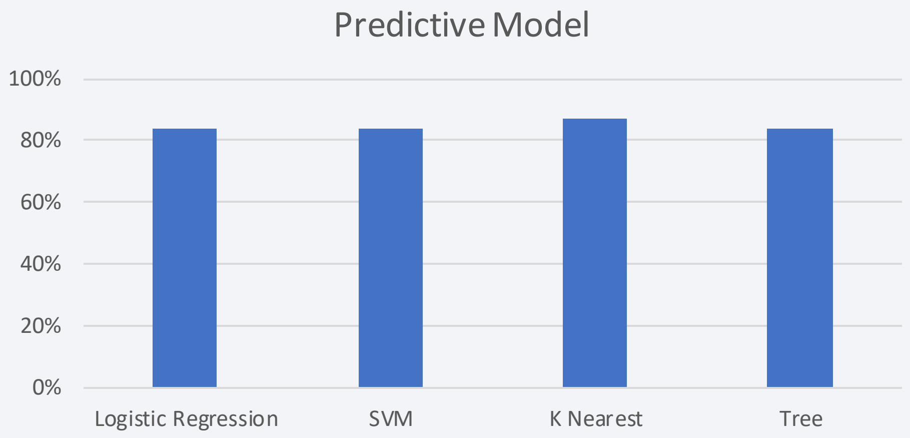
- Drone landings were most successful during 2010-2017.

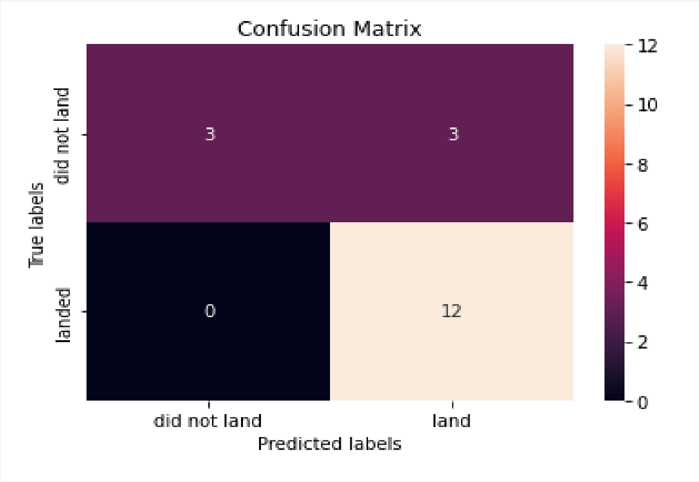
- Feature Selection: FlightNumber, PayloadMass, Orbit, LaunchSite. Flights, G



Results – Predictive Analysis

- The Logistic Regression, SVM and K Nearest Neighbor all had accuracy of 84% and precision of 80%.
- There is a risk that our prediction model will underestimate the number that did not land successfully.



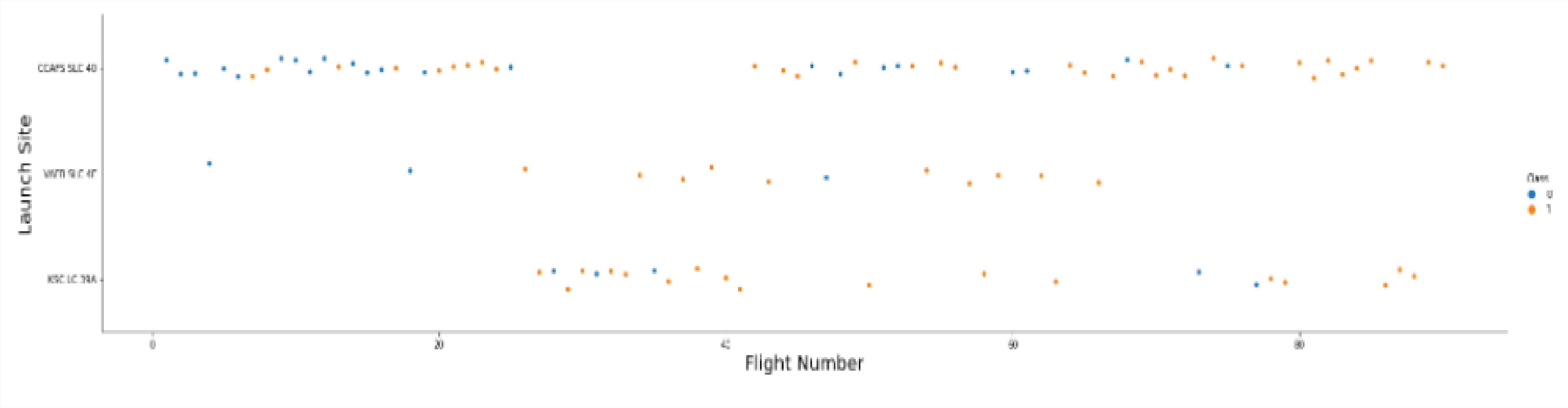


Section 2

Insights drawn from EDA

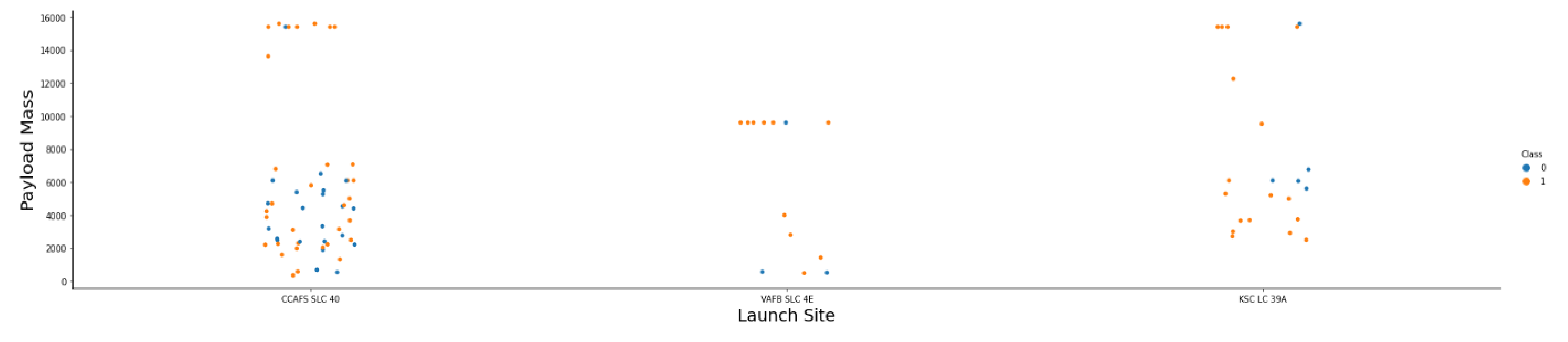
Flight Number vs. Launch Site

- CCAFS-SLC 40 Launch site has had the most launches and as flight attempts have increased, the success rate has gone up to around 9:1
- VAFB-SLC 4E has been consistent in its success rate at about 5:1



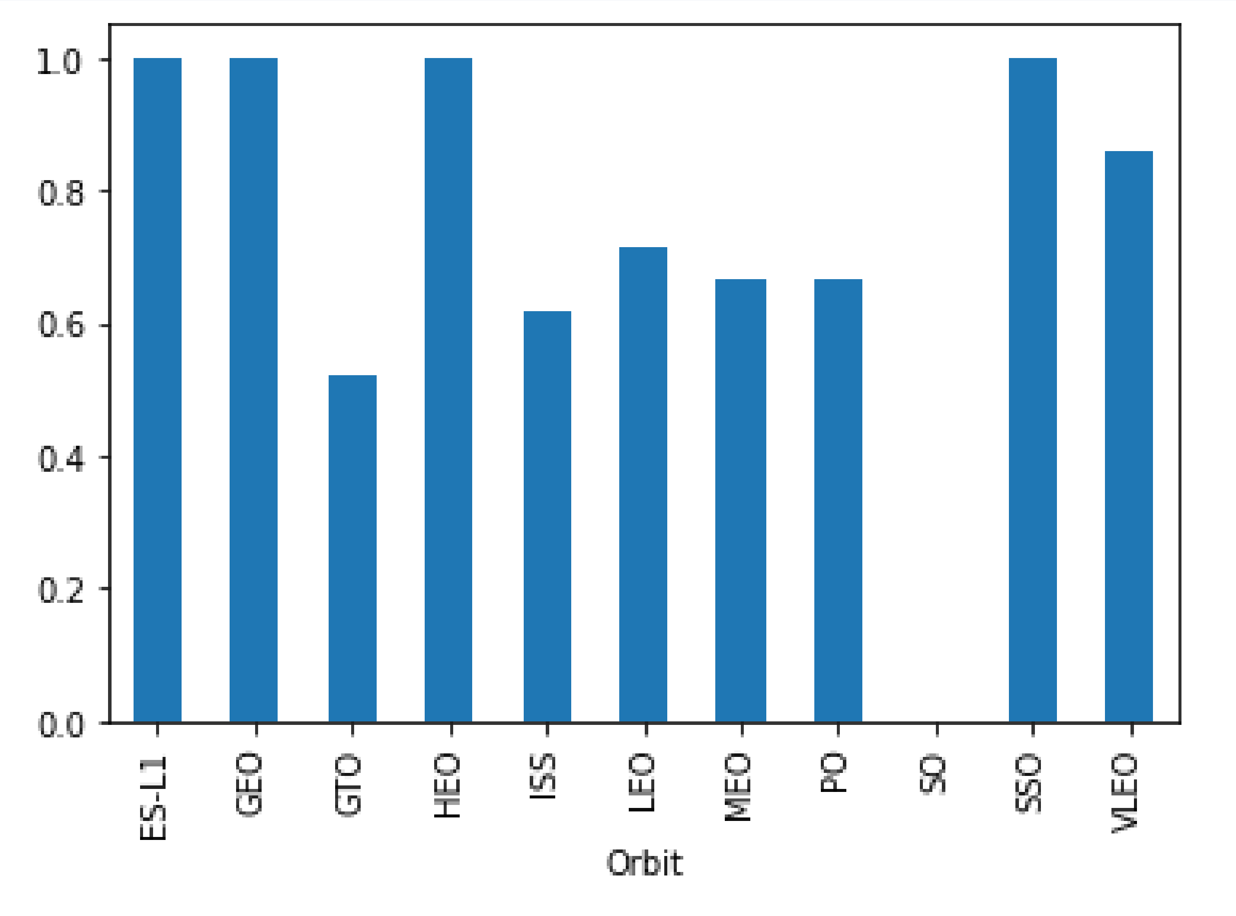
Payload vs. Launch Site

- CCAFS SLC 40 and KSC LC 39A has more success as Pay Load Mass goes up
- VAFB-SLC 4E launch site has no rockets launched for heavy payload mass(greater than 10,000KG)



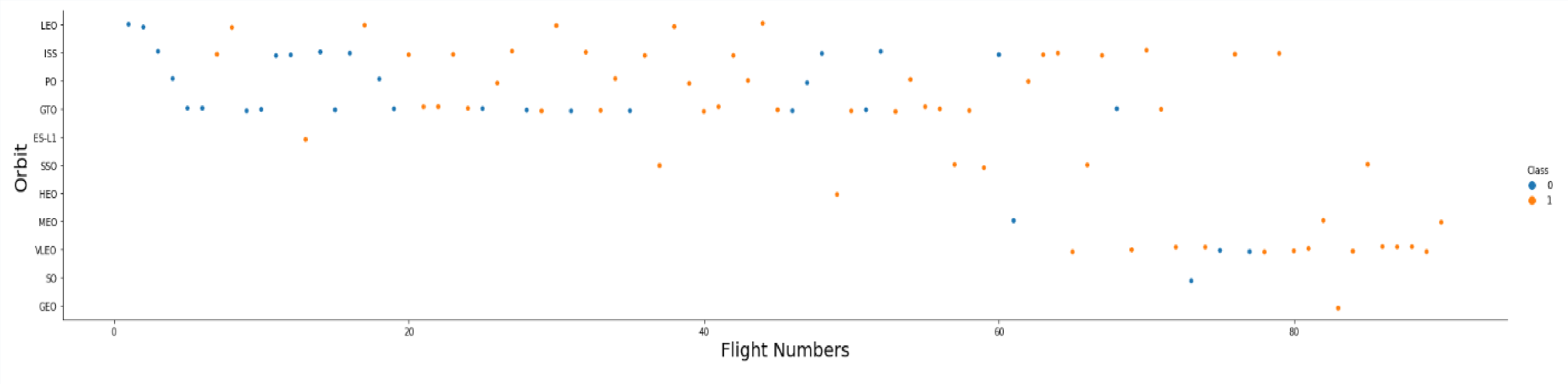
Success Type vs. Orbit Type

- GTO does not improve as payload mass increases
- Heavier payloads over 8,000 KG were only run in Polar, VLEO and ISS oribts.
- SSO orbit has a higher success rate but only with smaller payloads.



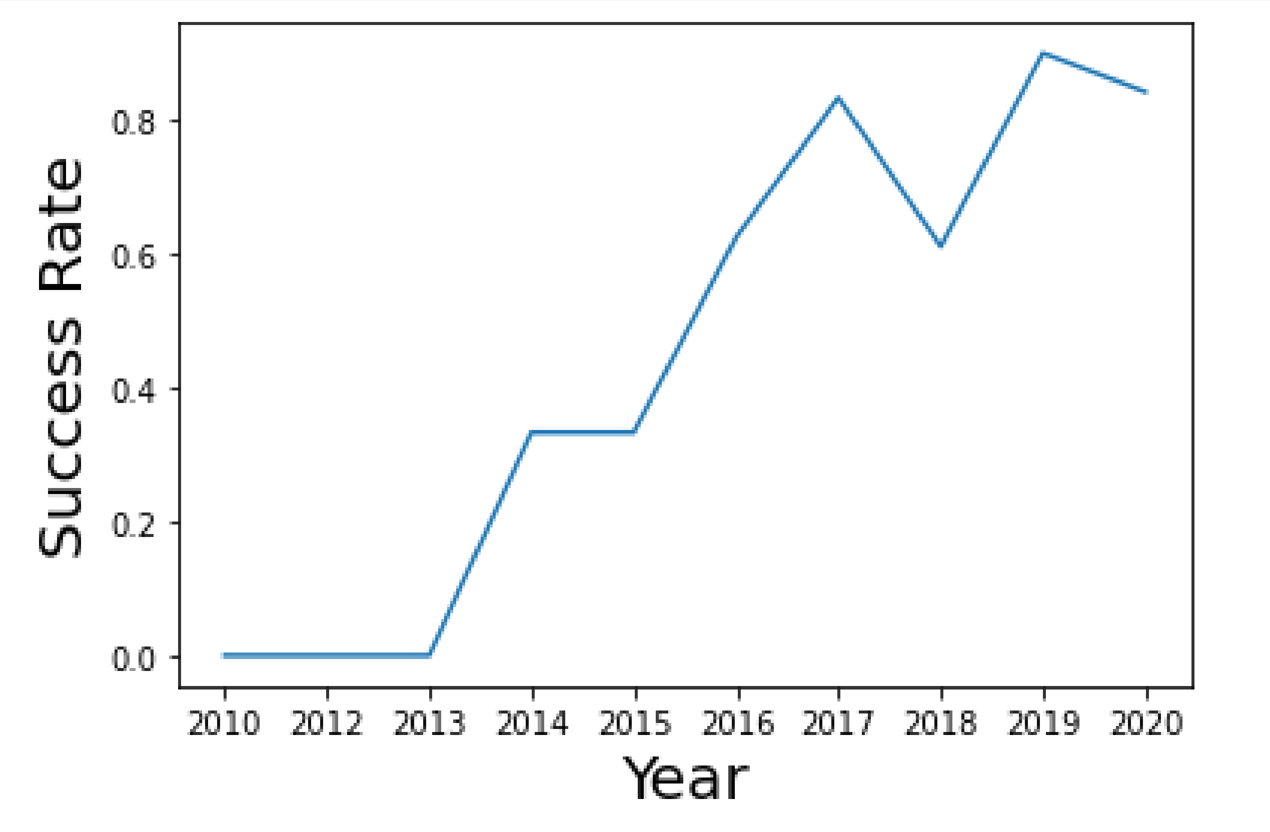
Flight Number vs. Orbit Type

- ISS, VLEO and LEO orbits seems to have a higher number of successful flights.
- There seems to be no relationship between the number of flights and GTO.
- All SSO flights were successful.



Launch Success Yearly Trend

- Average launch success rate has improved upwards since 2013.
- Seems a slight setback in 2018, but then the trend gets back on track.



All Launch Site Names

There are 4 unique launch site names from the web scrape data.

Out[7]:

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

Launch Site Names Begin with 'CCA'

The mission outcomes were all successful for the 5 `CCA` shown below in the LEO(ISS), GTO and LEO orbits

```
%sql select  * from Spacextbl 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

The total payload carried for NASA is 45,596 KG

```
%sql select  SUM(PAYLOAD_MASS__KG_) from Spacextbl WHERE CUSTOMER = 'NASA (CRS)';
```

1
45596

Average Payload Mass by F9 v1.1

- The average payload was queried from the dataset

```
%sql select  AVG(PAYLOAD_MASS__KG_) from Spacextbl WHERE BOOSTER_VERSION = 'F9 v1.1';
```

1
2928

First Successful Ground Landing Date

The first successful landing outcome on ground pad was December 22, 2015 for Spacex.

```
%sql select MIN(DATE) from Spacextbl WHERE LANDING__OUTCOME ='Success (ground pad)';
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

4 boosters have successfully landed on a drone ship and had payload mass greater than 4000 but less than 6000.

```
%sql select BOOSTER_VERSION from Spacextbl WHERE PAYLOAD_MASS__KG_ BETWEEN '4000' AND '6000' AND LANDING__OUTCOME ='Success (dro
ne ship)';
```

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

Total Number of Mission Outcomes

- There were 10 failed outcomes
- There were 69 successful 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

Boosters Carried Maximum Payload

There were 12 Boosters with the same maximum payload of 15,600 KG.

```
%sql select BOOSTER_VERSION from Spacextbl WHERE payload_mass__kg_=(select MAX(payload_mass__kg_) from Spacextbl);
```

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

2015 Launch Records

In 2015 there were 2 failed drone ship landings at Launch site CCAFS LC-40 by booster version F8 v1.1 B10XX .

```
%sql select BOOSTER_VERSION, DATE, LAUNCH_SITE from Spacextbl WHERE DATE BETWEEN '1-1-2015' and '1-1-2016' AND LANDING__OUTCOME = 'Failure (drone ship)';
```

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

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

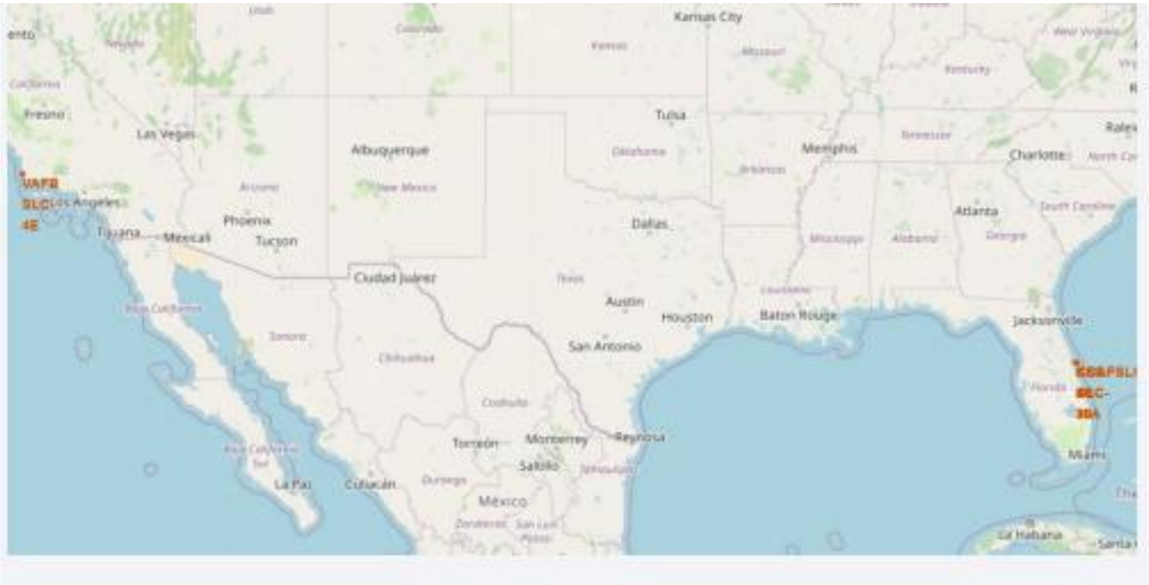
Between 2010 and 2017 SpaceX had a number of successes and failures. The ranking were pulled from the dataset.

```
%sql SELECT LANDING__OUTCOME,COUNT(LANDING__OUTCOME) FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' AND LANDING__OUTCOME LIKE 'Success%'GROUP BY LANDING__OUTCOME
```

landing__outcome	2
Success (drone ship)	5
Success (ground pad)	3

Section 3

Launch Sites Proximities Analysis

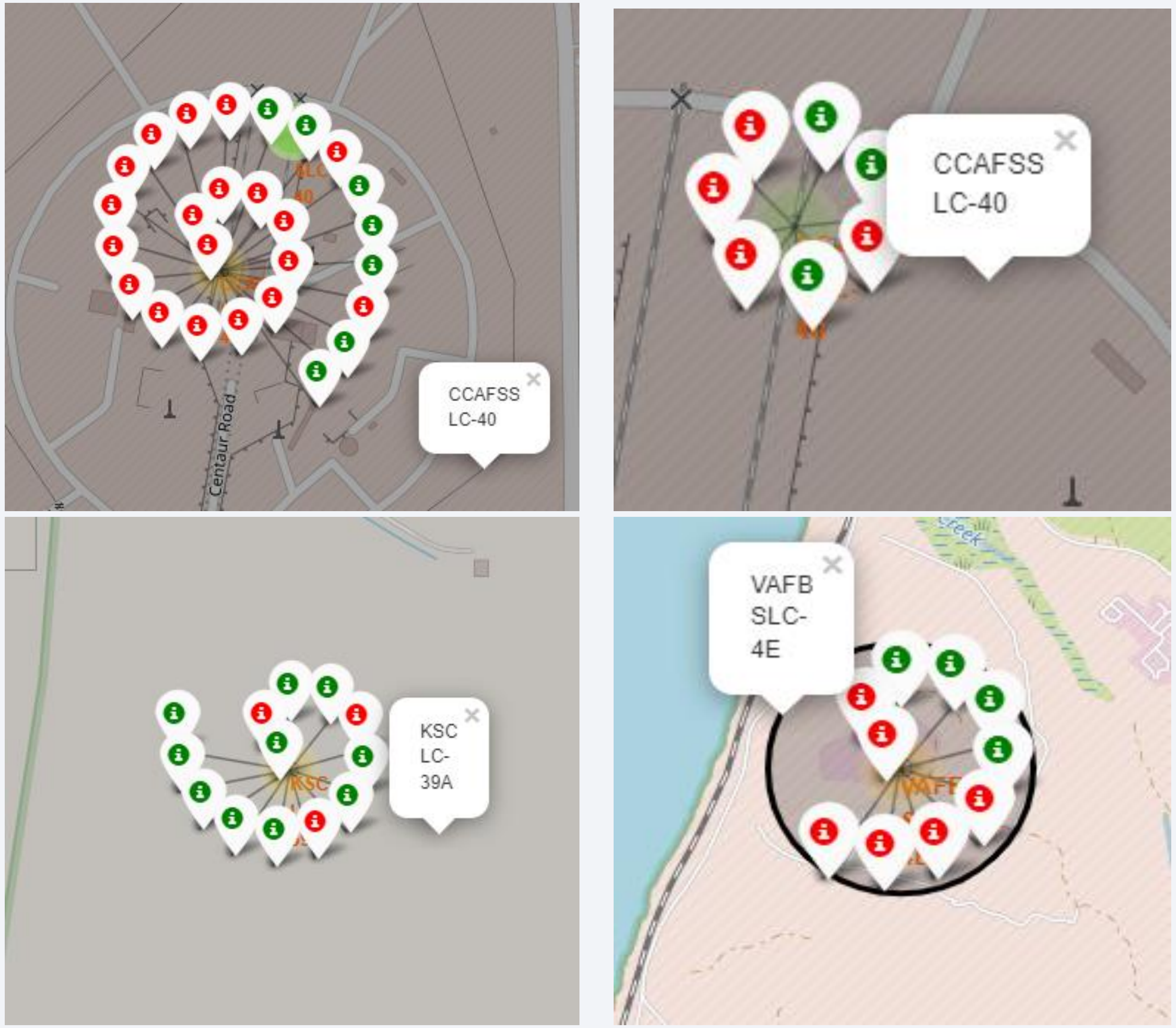


Launch Site Locations

- CCAFS SLC-40 and CCAFS LC-40 the same place.
- There are only 3 launch sites used and they are all near the ocean
- Launch sites are usually near the equator because of the rotation of the earth gives the launch extra speed.

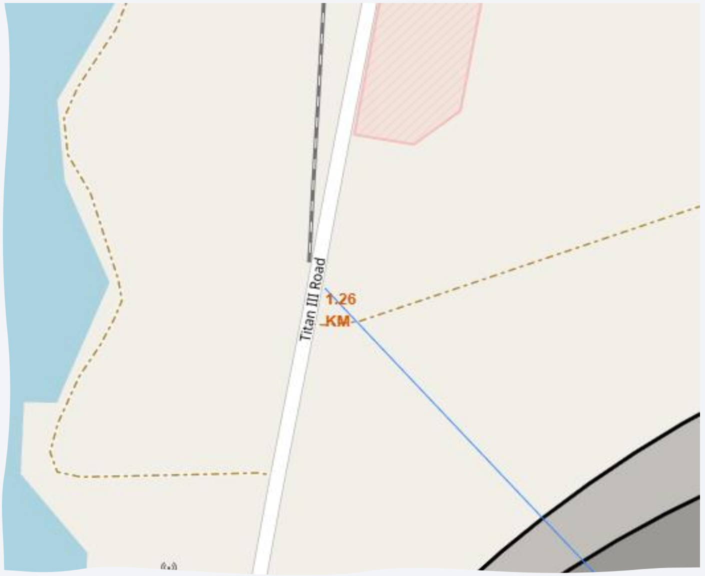
Launch Outcomes

- SLC LC 39A had the greatest success rate
- Red icons are failed launches and green are successful



Launch Site Proximity

- CCAFS SLC-40 site is closest to the ocean at 0.85 km
- CCAFS SLC-40 is also 1.25 km from the Nasa Railroad for easier transport to the launch site.

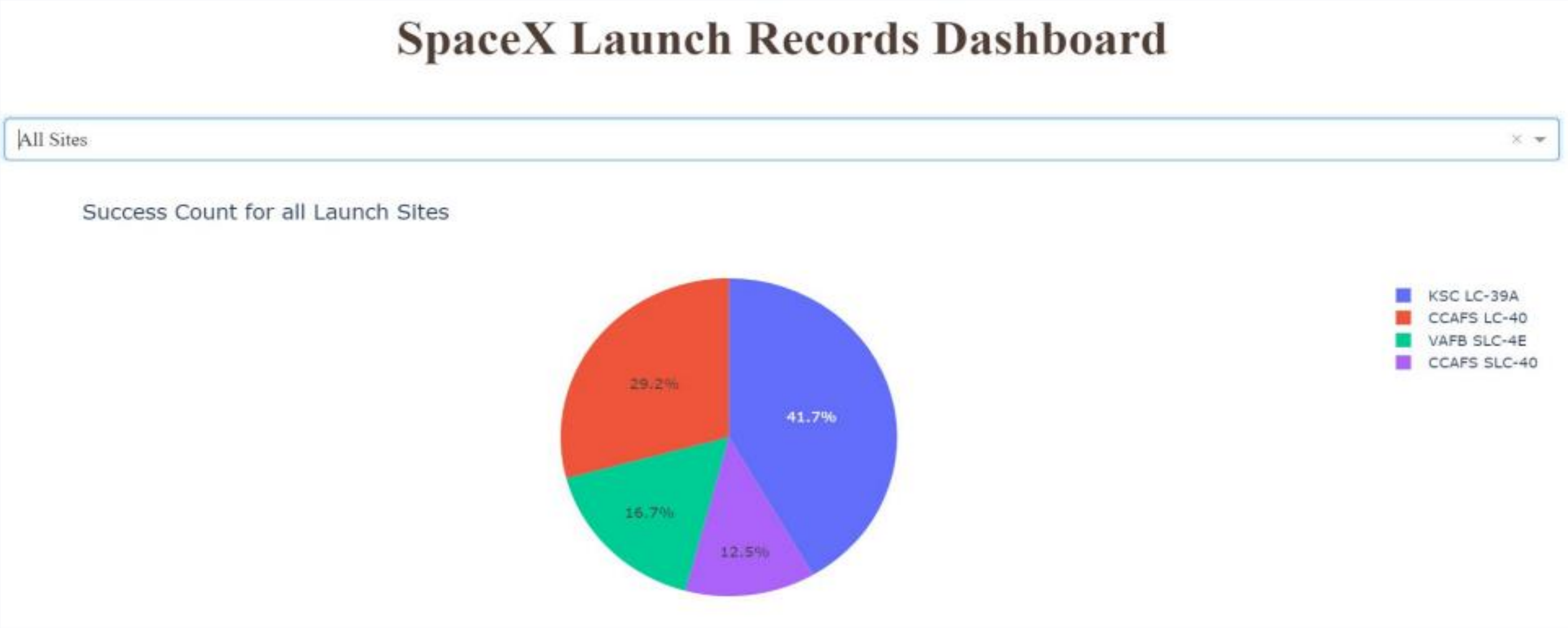


Section 4

Build a Dashboard with Plotly Dash

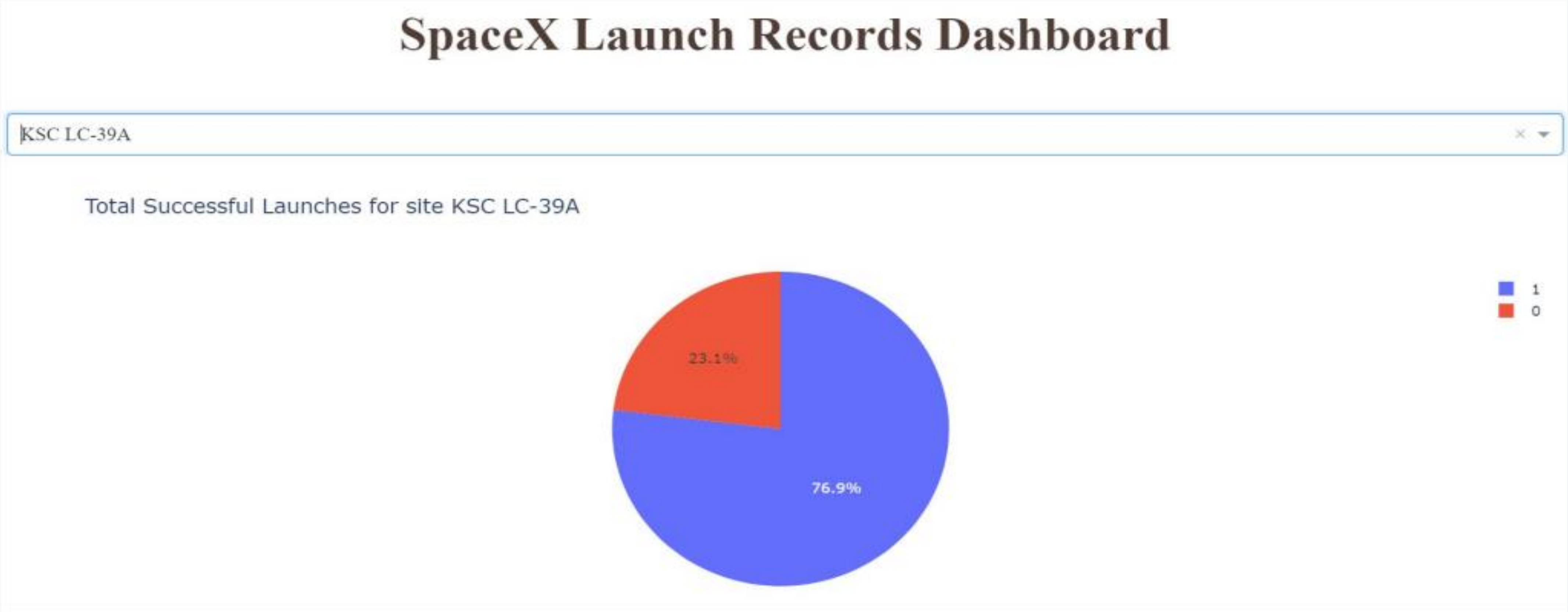
Launch Success by Site

A dashboard was created for further EDA on the dataset



Highest Launch Success Rates

KSC LC-39A has one of the highest success rates



Success Rates of Different Boosters

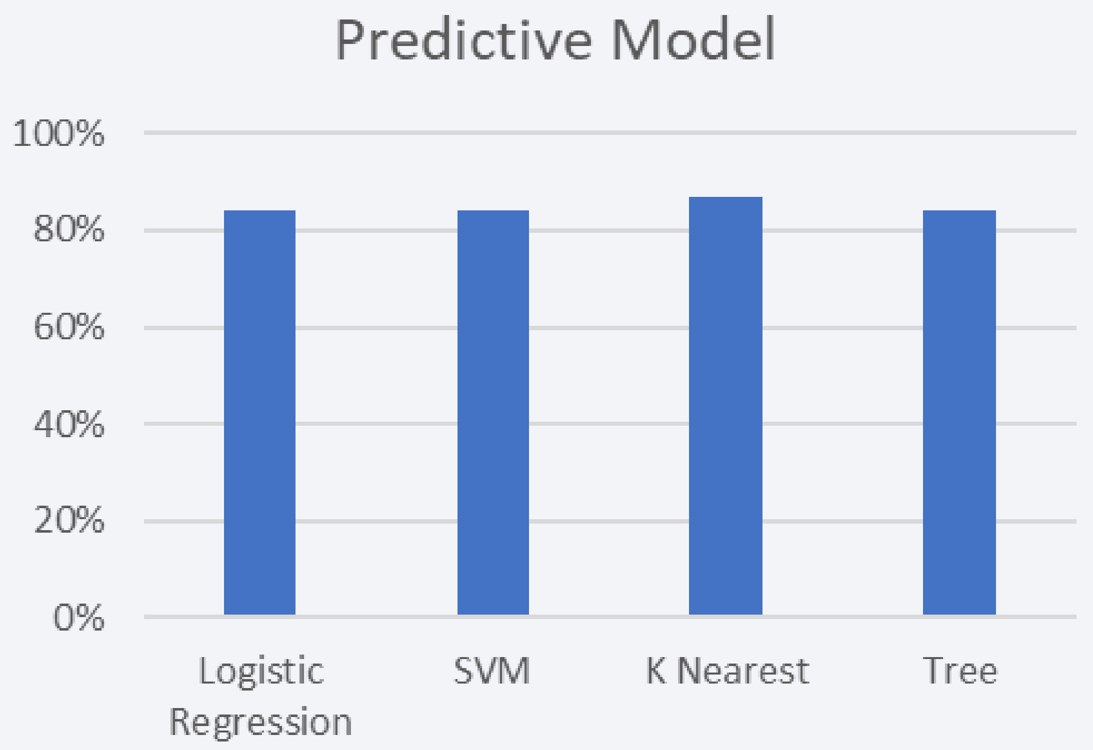
FT appears to have the highest success rate at higher Payload Masses.

Section 5

Predictive Analysis (Classification)

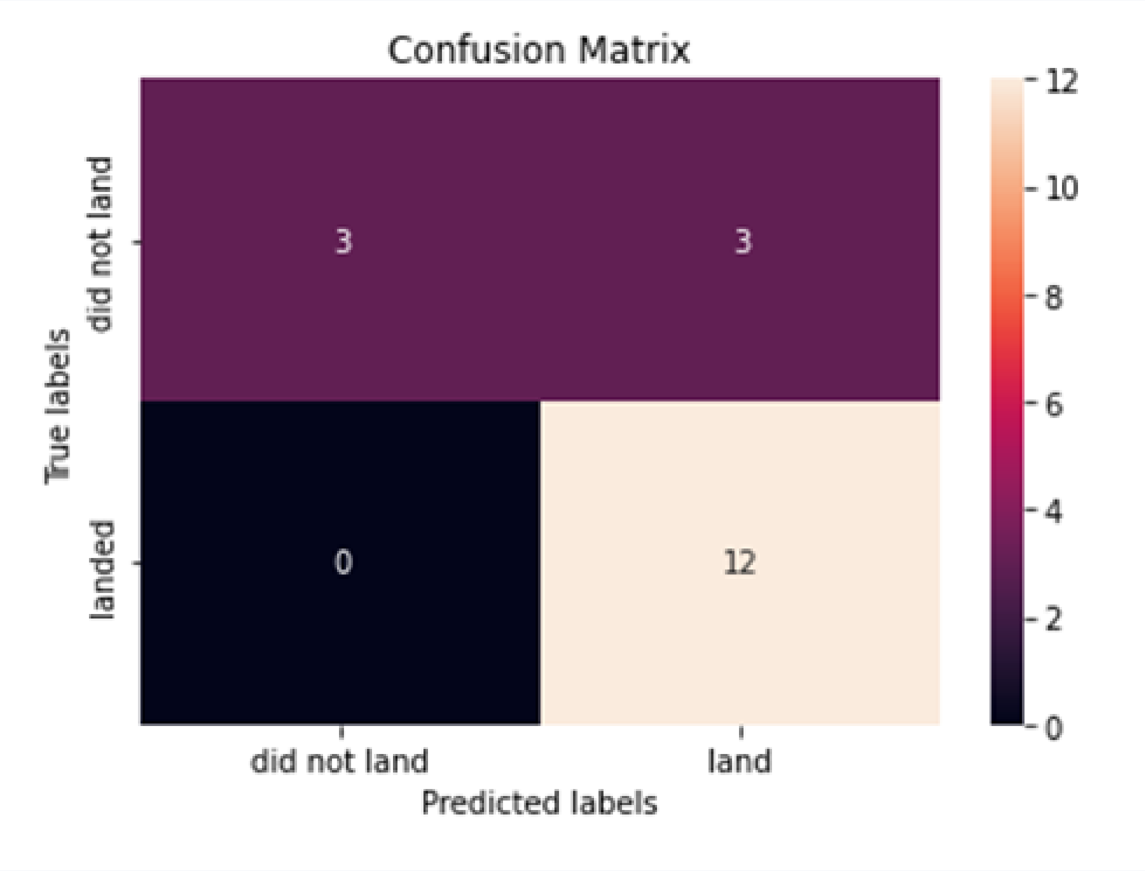
Classification Accuracy

Logistic Regression, SVM , K Nearest Neighbor and Tree Classifications all had about the same accuracy



Confusion Matrix

- The confusion matrix was the same for all prediction models.
- A larger dataset would probably separate these prediction models more.
- Models seem very good at predicting landings but not crashes



Conclusion

- Accuracy could be improved with the current dataset.
- Not being able to accurately predict crashes could be too high a risk for Star Aeronautics.
- Star Aeronautics should continue improving the launch models using the items listed below.

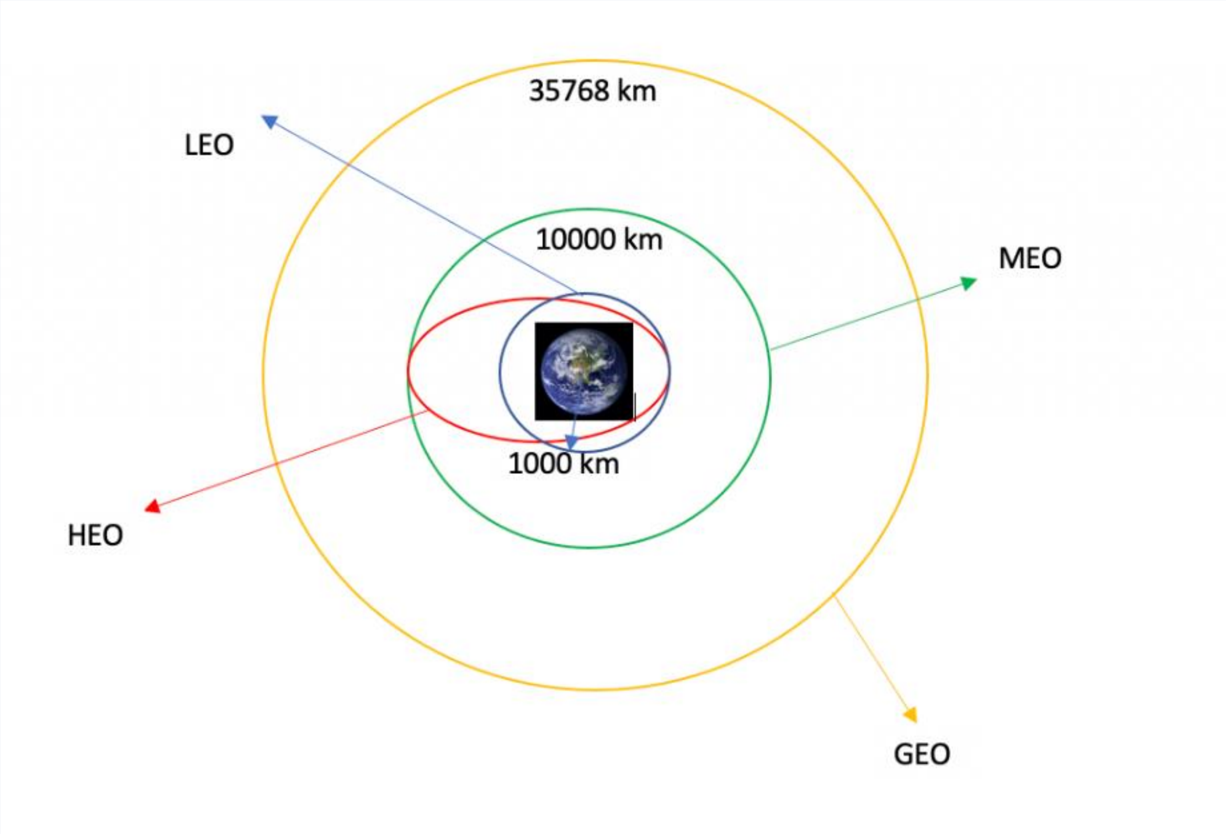
1. Adding more data
2. Treat missing and outliner values differently
3. Feature Engineering
4. Selecting additional features
5. Model tuning

Appendix - Definitions

- Payload: Purpose of mission. Could include satellites, telescopes, Space Station Modules, Cargo, and even crew spacecraft
- Staging: Staging is the combination of several rocket sections, or stages, that fire in a specific order and then detach, so a ship can penetrate Earth’s atmosphere and reach space
- Fins: Grid fins are a type of flight control surface used on rockets and bombs.
- Legs: Enables the phases of the rocket to be land and be reusable.
- Data Wrangling: Data wrangling is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
- Outcome: Whether first stage successfully landed True ASDS / Landed to drone ship, False ASDS / Didn’t land to drone ship.

Appendix - Orbit

Orbit: An orbit is a regular, repeating path that one object in space takes around another one. There are three main "bands" of orbit around the Earth: low Earth orbit (LEO), medium Earth orbit (MEO), and geostationary orbit (GEO).



Appendix – Orbit cont.

Orbits

- LEO: Low Earth orbit (LEO)is an Earth-centered orbit with an altitude of 2,000 km (1,200 mi) or less (approximately one-third of the radius of Earth),[1] or with at least 11.25 periods per day (an orbital period of 128 minutes or less) and an eccentricity less than 0.25.[2] Most of the manmade objects in outer space are in LEO.
- VLEO: Very Low Earth Orbits (VLEO) can be defined as the orbits with a mean altitude below 450 km. Operating in these orbits can provide a number of benefits to Earth observation spacecraft as the spacecraft operates closer to the observation.
- GTO A geosynchronous orbit is a high Earth orbit that allows satellites to match Earth's rotation. Located at 22,236 miles (35,786 kilometers) above Earth's equator, this position is a valuable spot for monitoring weather, communications and surveillance. Because the satellite orbits at the same speed that the Earth is turning, the satellite seems to stay in place over a single longitude, though it may drift north to south," NASA wrote on its Earth Observatory website.
- SSO (or SO): It is a Sun-synchronous orbit also called a Helio synchronous orbit is a nearly polar orbit around a planet, in which the satellite passes over any given point of the planet's surface at the same local mean solar time.

Appendix – Orbit cont.

Orbits

- ES-L1 :At the Lagrange points the gravitational forces of the two large bodies cancel out in such a way that a small object placed in orbit there is in equilibrium relative to the center of mass of the large bodies. L1 is one such point between the sun and the earth.
- HEO A highly elliptical orbit, is an elliptic orbit with high eccentricity, usually referring to one around Earth.
- ISS A modular space station (habitable artificial satellite) in low Earth orbit. It is a multinational collaborative project between five participating space agencies: NASA (United States), Roscosmos (Russia), JAXA (Japan), ESA (Europe), and CSA (Canada)
- MEO Geocentric orbits ranging in altitude from 2,000 km (1,200 mi) to just below geosynchronous orbit at 35,786 kilometers (22,236 mi). Also known as an intermediate circular orbit. These are "most commonly at 20,200 kilometers (12,600 mi), or 20,650 kilometers (12,830 mi), with an orbital period of 12 hours
- HEO Geocentric orbits above the altitude of geosynchronous orbit (35,786 km or 22,236 mi)
- GEO It is a circular geosynchronous orbit 35,786 kilometers (22,236 miles) above Earth's equator and following the direction of Earth's rotation PO It is one type of satellites in which a satellite passes above or nearly above both poles of the body being orbited (usually a planet such as the Earth

Appendix – links to images used

- https://www.smallcapnews.co.uk/spacexs-cargo-dragon-2s-first-recovery-was-delayed-due-to-atlantic-weather/
- https://i1.wp.com/www.healththoroughfare.com/wp-content/uploads/2020/03/spacex-dragon-1-capsule-cargo-delivery.jpg?fit=1200%2C800&ssl=1
- https://external-content.duckduckgo.com/iu/?u=https%3A%2F%2Ftse2.mm.bing.net%2Fth%3Fid%3DOIP.yrRKOGV0z9bbfNSAFvEg3AHaD3%26pid%3DApi&f=1
- https://external-content.duckduckgo.com/iu/?u=https%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2Fc510b19f92d7c588e80e3f400de6570d4e9e8083%2Fc%3D0-295-1998-1424%2Flocal%2F-%2Fmedia%2F2018%2F02%2F19%2FBrevard%2FBrevard%2F636546462900293572-372A8212.jpg%3Fwidth%3D1998%26height%3D1129%26fit%3Dcrop%26format%3Dpjpg%26auto%3Dwebp&f=1&nofb=1
- https://external-content.duckduckgo.com/iu/?u=https%3A%2F%2Ftse1.mm.bing.net%2Fth%3Fid%3DOIP.ZyYMjKQULxeiQKPwO1qz7gHaE8%26pid%3DApi&f=1
- https://external-content.duckduckgo.com/iu/?u=https%3A%2F%2Ftse3.mm.bing.net%2Fth%3Fid%3DOIP.-utG70tl5uOqJxQqqPa6rQHaEK%26pid%3DApi&f=1
- https://cdnph.upi.com/svc/sv/upi/6951536350517/2018/1/1de0476983ea6ee11f46f93cff6096b3/SpaceXs-Falcon-9-launches-powerful-telecom-satellite-into-orbit.jpg
- https://www.theregreview.org/wp-content/uploads/2020/06/regulatingcommercialspace-768x480.jpg

Thank you!

- Thank you to Star Aeronautics for the opportunity
- Big Thanks to IBM for their data science tools and services
- Thank you to Rav Ahuja for the excellent assistance