IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Hasnain Naeem
15th November 2022

Outline

Executive Summary

- In this report, we predict if the Falcon 9 first stage will land successfully or not so that our company SpaceY can determine the cost of the launch. In order to get the prediction, we extracted data through SpaceX API and webscraping a Wikipedia page using Beautiful Soup. Upon extraction, we processed the data by removing unwanted columns, handling null values and assigning codes to categorical values.
- On the processed data, we performed exploratory data analysis through SQL and visualization. Through this we understood the relations between different variables and shortlisted the variables that we needed to use in our machine learning algorithms.
- We divided our dataset into training and testing set and ran multiple algorithms (KNN, SVM, Logistical Regression and Tree) on the training data set and tested its accuracy on the testing set. To get the best hyperparameters, we used cross validation. In the end, we discovered that all the model used had equal accuracy and we were able to successfully predict the landing at an accuracy of 83%.

Introduction

- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost more than 165 million dollars each.
- The reason for the low cost for SpaceX is due to their technology to reuse the first stage.
- However, the first stage is not always landed successfully.
- In this report, we will predict if the Falcon 9 first stage will land successfully or not so that our company SpaceY can determine the cost of the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Request to the SpaceX API
  - Web scraping to collect Falcon 9 historical launch records from a Wikipedia page
- Perform data wrangling
  - Conversion of JSON to dataframe
  - Filtering required data
  - Null values were replaced with mean values

Methodology (continued)

Executive Summary

- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - Identified variables to be used by models by looking at relationship between variables
    - Divided the data into training and test sets
    - Ran multiple algorithms(Logistical Regression, KNN, SVM and Tree) and performed cross validation
    - Checked accuracy score and confusion matrix for evaluating the model

Data Collection

Two methods to collect the required data:

- Request to the SpaceX API
  Rocket launch data including **flight number, data, booster version, payload mass, orbit, launch site, outcome, landing pads, cordinates, grid fins etc.**

- Web scraping to collect Falcon 9 historical launch records from a Wikipedia page titled **List of Falcon 9 and Falcon Heavy Launches**
  Beautiful soup was used to collect rocket launch data including **flight number, data, booster version, payload mass, orbit, launch site, outcome, landing pads, cordinates, grid fins etc.**

Data Collection – SpaceX API

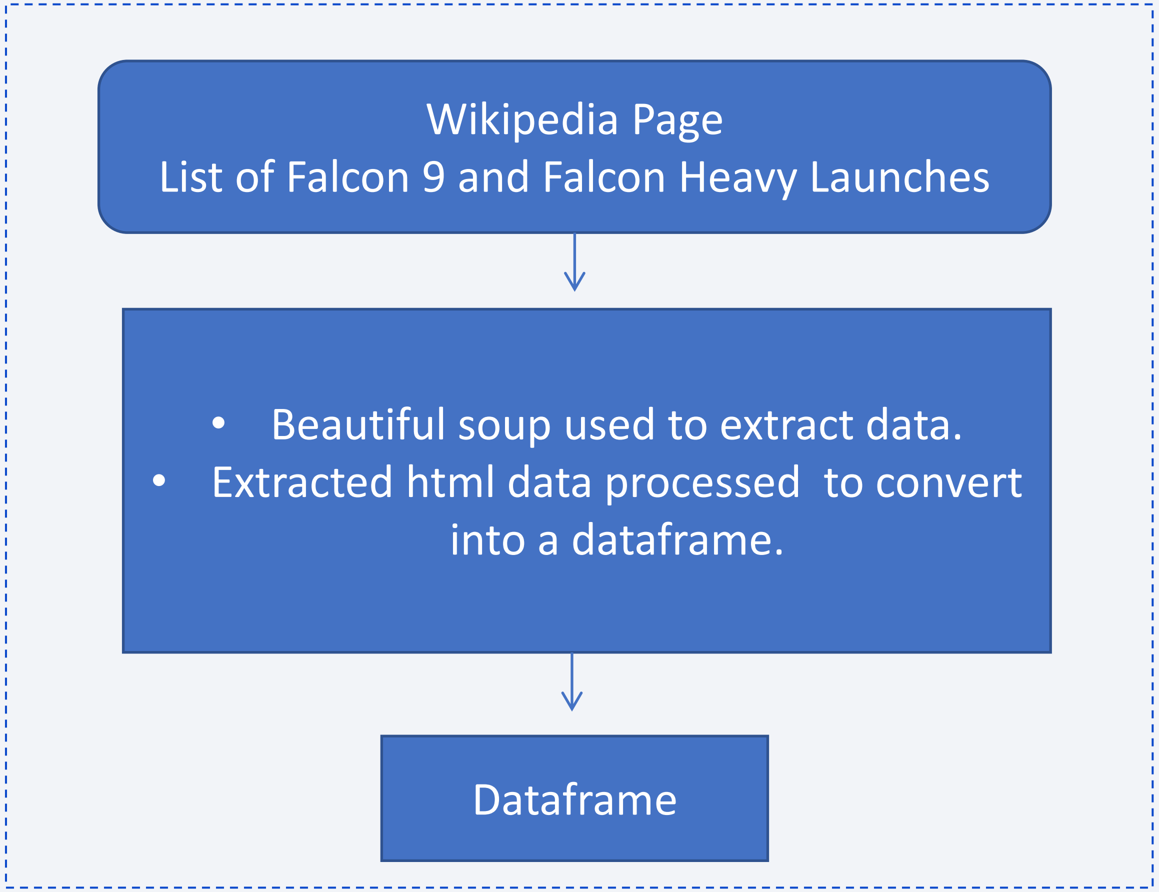
- Request to the SpaceX API to get the following information as JSON object in response
  Flight number, data, booster version, payload mass, orbit, launch site, outcome, landing pads, cordinates, grid fins etc.

- https://github.com/hasnainnaeem93/IBMFinalProject/blob/master/Falcon9_DataCollection.ipynb

Data Collection - Scraping

- Web scraping using Beautiful Soup package was done to collect Falcon 9 historical launch records from a Wikipedia page titled **List of Falcon 9 and Falcon Heavy Launches** to collect rocket launch data including **flight number, data, booster version, payload mass, orbit, launch site, outcome, landing pads, cordinates, grid fins etc.**

- https://github.com/hasnainnaeem93/IBMFinalProject/blob/master/Falcon9_WebScraping.ipynb



Data Wrangling

- Data was studied and understood.
- Counts of different orbits, launch sites and outcomes were looked.
- Missing values were replaced by mean.
- The outcomes were categorized as success or failure and then added as 0 or 1 for ease in analysis and applying predictive models.
- https://github.com/hasnainnaeem93/IBMFinalProject/blob/master/Falcon9_EDA.ipynb

EDA with Data Visualization

- Scatter plots, bar chart and line chart were used for visualization of exploratory data analysis to understand the relation between different parameters and trends over a time period.
- https://github.com/hasnainnaeem93/IBMFinalProject/blob/master/Falcon9_EDA_Visualization.ipynb

EDA with SQL

SQL queries were performed to extract:

- Distinct launch sites
- 5 records with launch sites starting with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1
- Date of first successful landing outcome in ground pad
- Distinct boosters which have success in drone ship and have payload mass between 4000 and 6000
- Number of successful and failed mission outcomes
- Booster versions which carried maximum payload
- Failed landing outcomes in drone ship and their booster versions for year 2015
- Landing outcomes and their counts

https://github.com/hasnainnaeem93/IBMFinalProject/blob/master/Falcon9_EDA_SQL.ipynb

Build an Interactive Map with Folium

Circles, markers and lines were added to the Folium map

- Circles were added to highlight different launch sites.
- Colored markers were added to show the type of outcome.
- The were added in a cluster for better readability.
- Lines were drawn to show distance between different points on the map e.g. the launch site and the coastline etc.

https://github.com/hasnainnaeem93/IBMFinalProject/blob/master/Falcon9_Visualization.ipynb

Build a Dashboard with Plotly Dash

- A pie chart and scatter plot have been added to the dashboard to understand the relation between launch sites, payload and outcomes.
- A dropdown list and slider have been added to make the dashboard interactive.
- https://github.com/hasnainnaeem93/IBMFinalProject/blob/master/Dash.py

Predictive Analysis (Classification)

- Performed EDA and determined training Labels
- Created a column for the class in-order to train models
- Standardized the data
- Split the dataframe into training data and test data
- Ran the following algorithms and dermined the best hyperparameter for each SVM, Classification Trees, KNN and Logistic Regression through Grid Search Cross Validation
- Ran each algorithm on test data to figure out the best performing model
- https://github.com/hasnainnaeem93/IBMFinalProject/blob/master/Falcon9_MachineLearning.ipynb

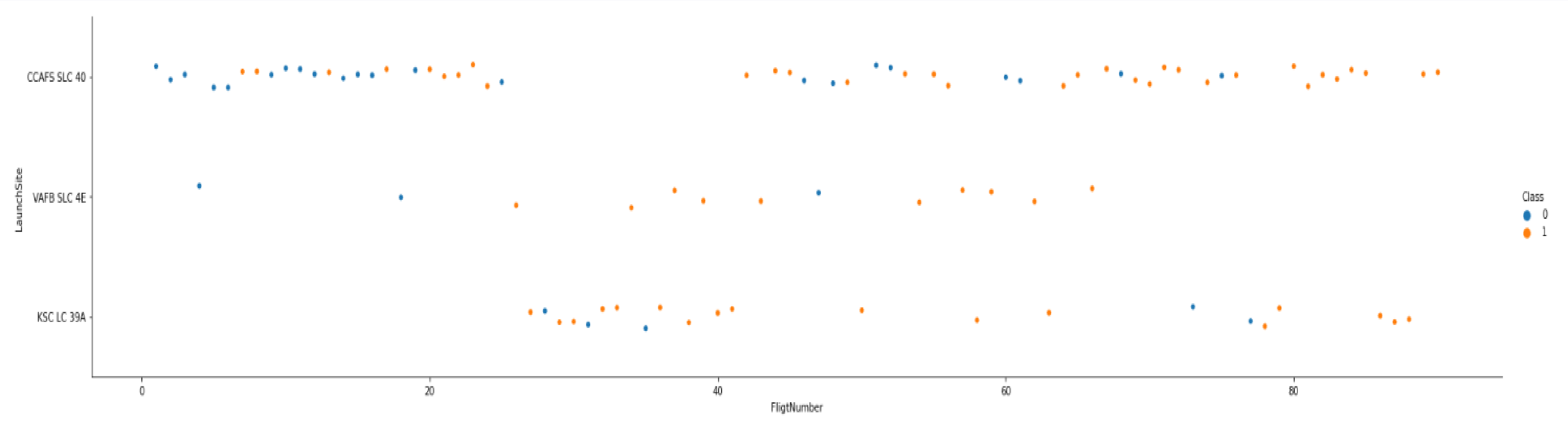
Results

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

Section 2

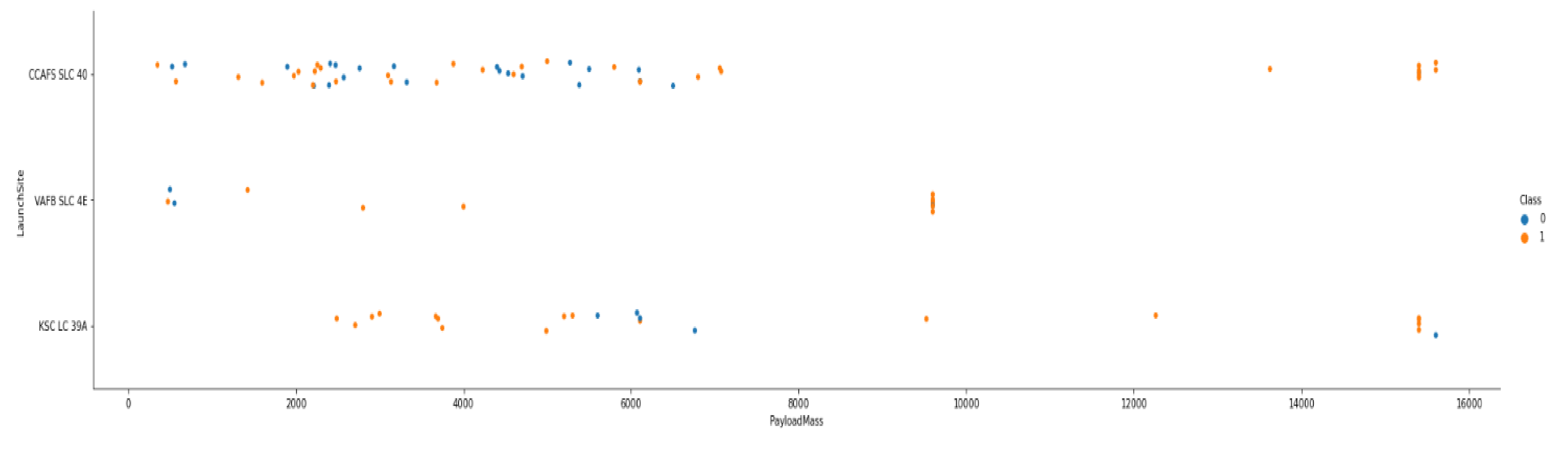
Insights drawn from EDA

Flight Number vs. Launch Site



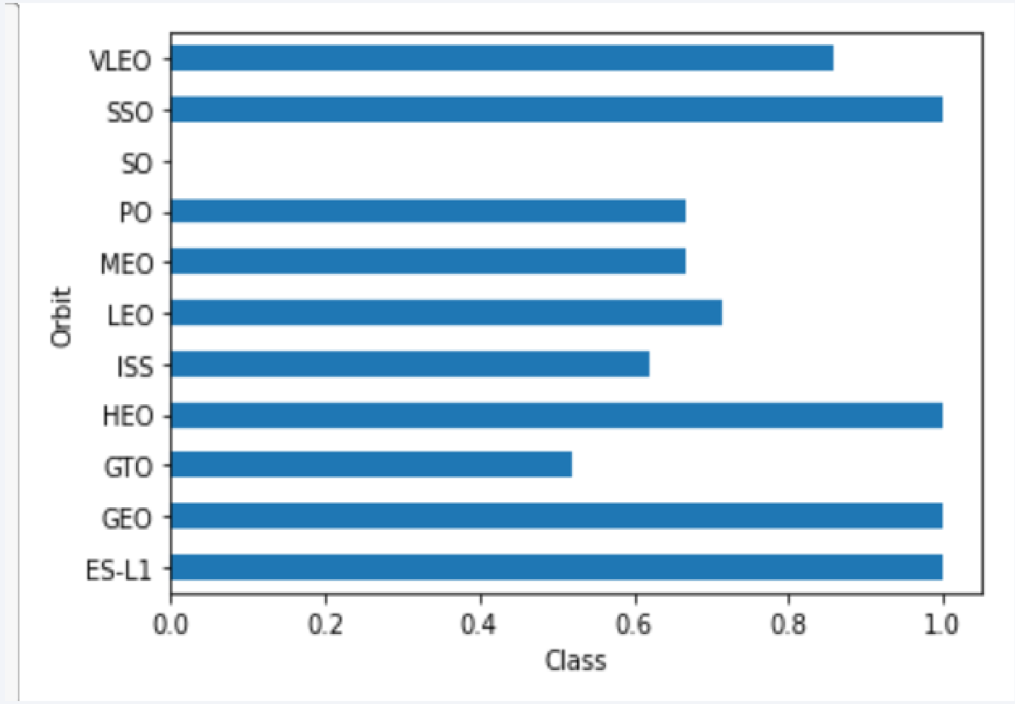
As the flight number increases the number of successful landings increase

Payload vs. Launch Site



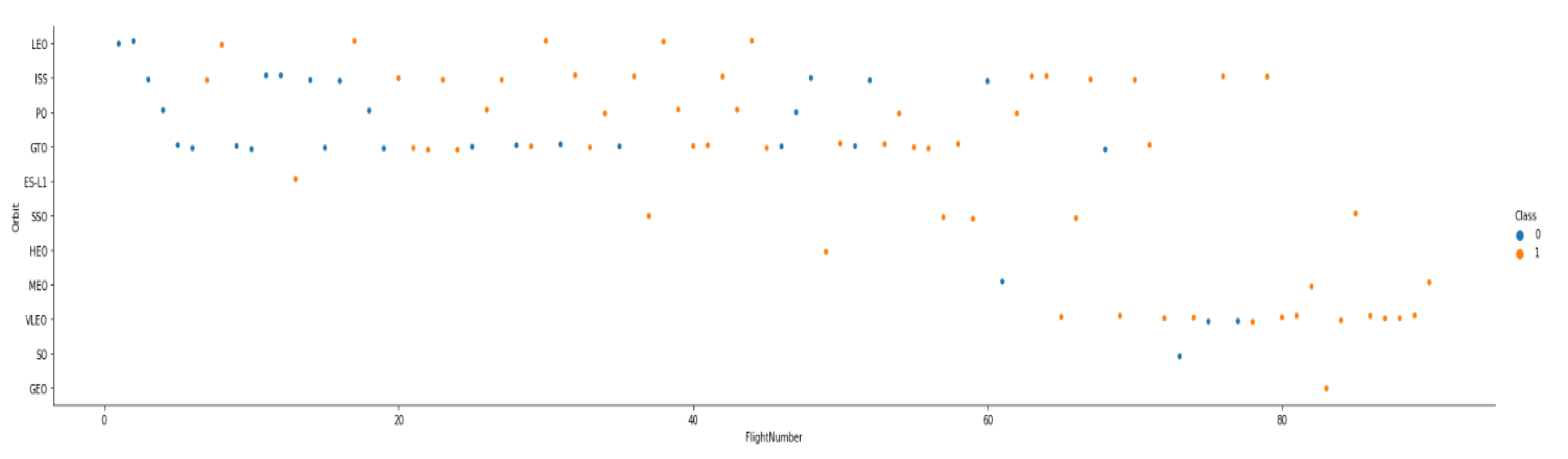
- For all launch sites, heavy payloads have a greater success rate for all launch sites.
- For VAFB-SLB, there are no payloads greater than 10000.

Success Rate vs. Orbit Type



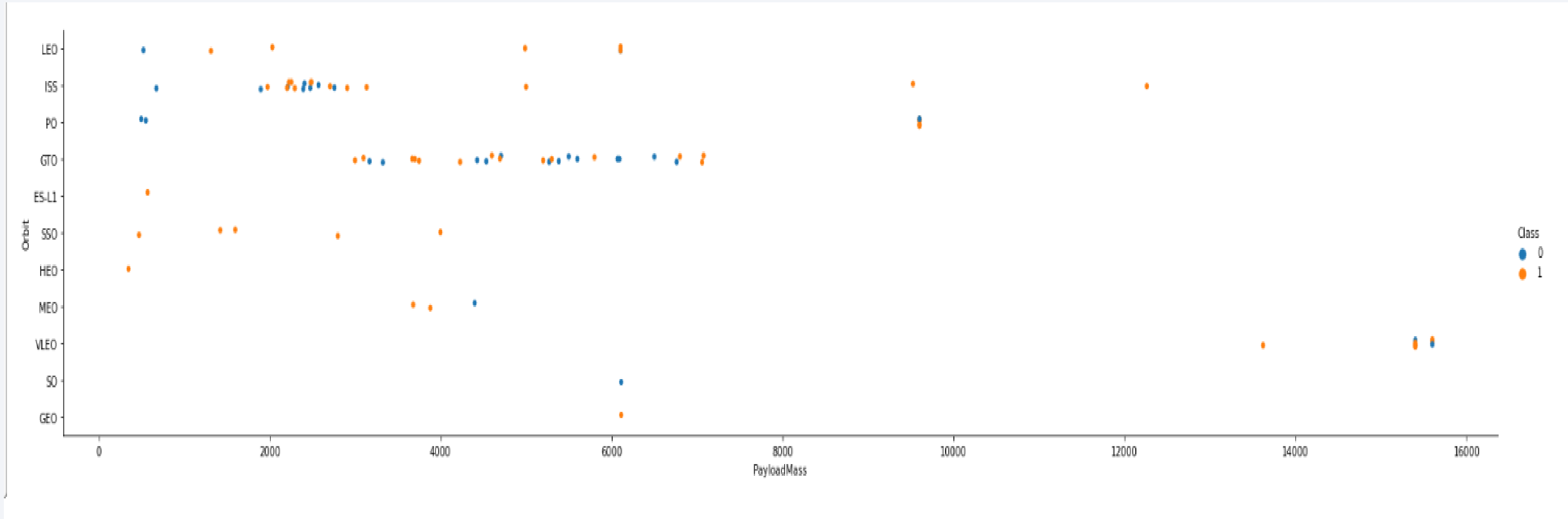
- SSO, HEO, GEO and ES-L1 orbits have a 100% success rate.

Flight Number vs. Orbit Type



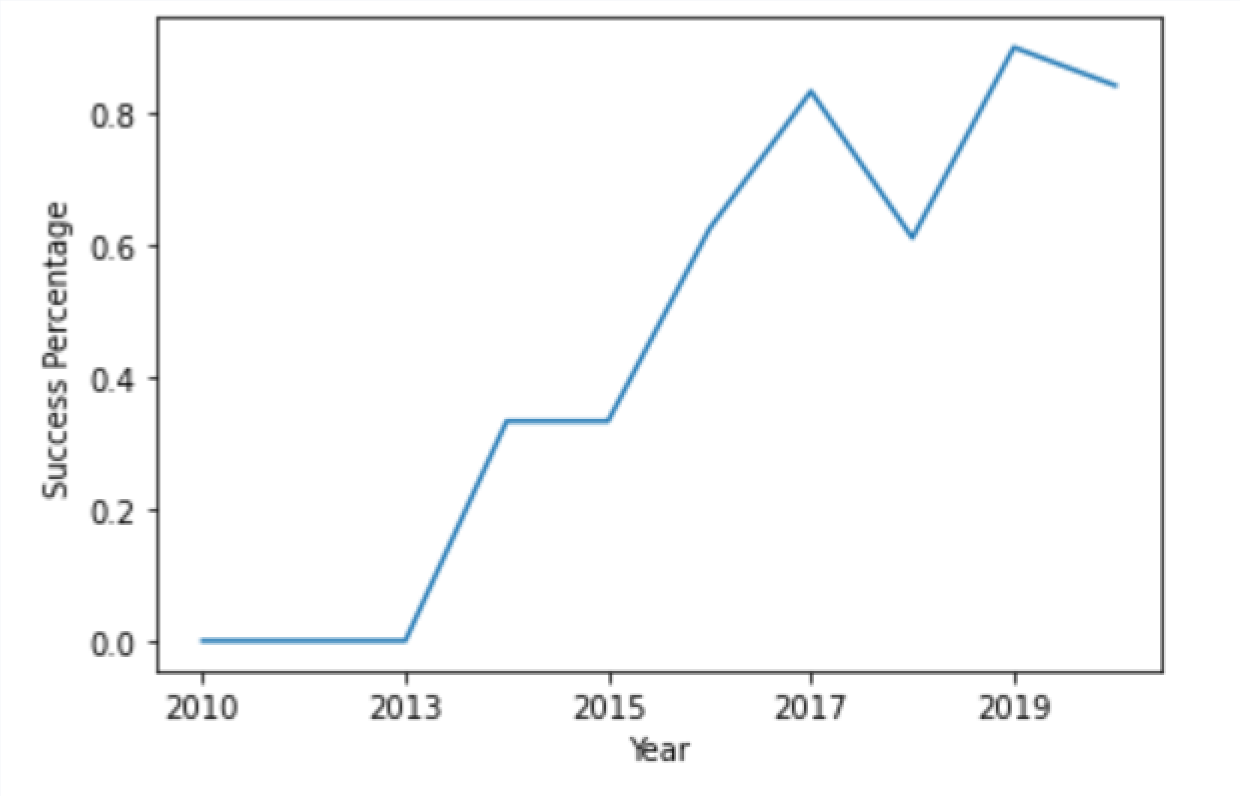
- In the LEO orbit the success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.

Payload vs. Orbit Type



- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here

Launch Success Yearly Trend



- The sucess rate since 2013 kept increasing till 2020.

All Launch Site Names

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

Launch Site Names Begin with 'CCA'

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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)
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-12	22:41:00	F9 v1.1	CCAFS LC-40	SES-8	3170	GTO	SES	Success	No attempt

Total Payload Mass

customer	total_payload_mass
NASA (CRS)	22007

Average Payload Mass by F9 v1.1

booster_version	avergae_payload_mass
F9 v1.1	3676

First Successful Ground Landing Date

minimum_date
2017-01-05

Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version
F9 FT B1031.2
F9 FT B1022

Total Number of Successful and Failure Mission Outcomes

mission_outcome	COUNT
Success	44
Success (payload status unclear)	1

Boosters Carried Maximum Payload

booster_version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1058.3
F9 B5 B1060.2

2015 Launch Records

landing__outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40

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

landing__outcome	COUNT
No attempt	7
Failure (drone ship)	2
Success (drone ship)	2
Success (ground pad)	2
Controlled (ocean)	1
Failure (parachute)	1

Section 3

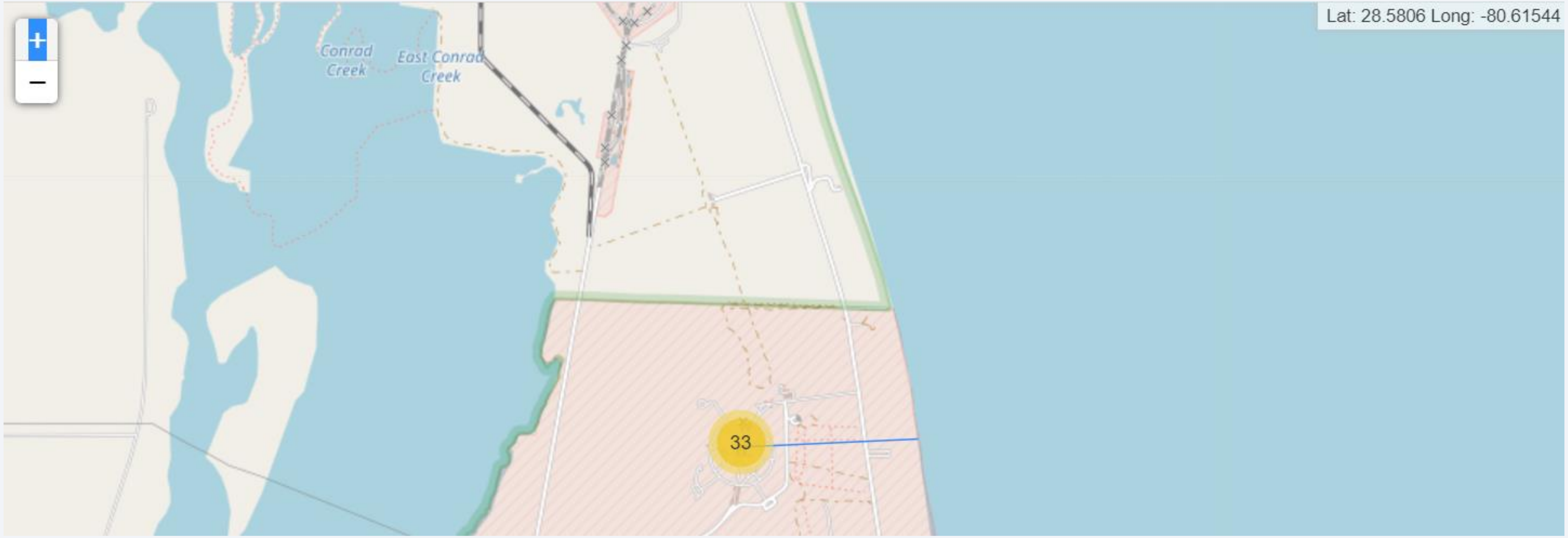
Launch Sites Proximities Analysis

All Launch Sites

Outcomes at Launch Sites

A cluster of all successful and unsuccessful landings at each launch site

Distance from Launch Site

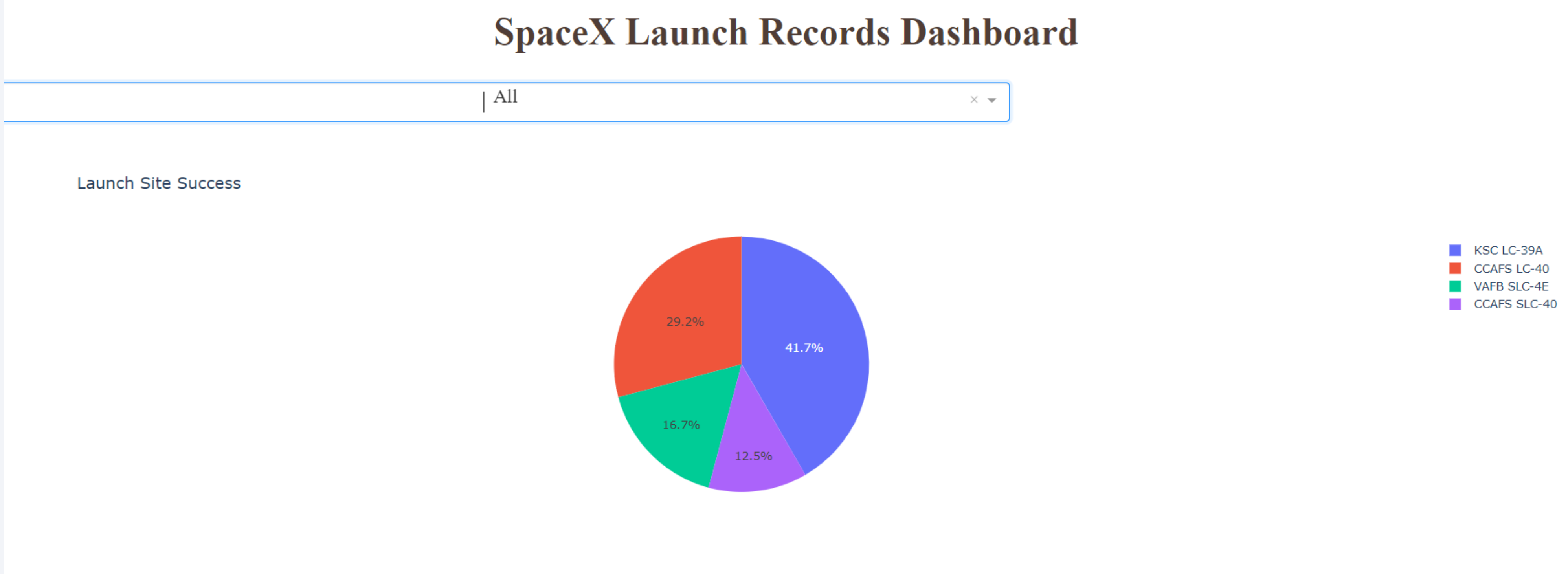


Distance of a launch site to the coastline

Section 4

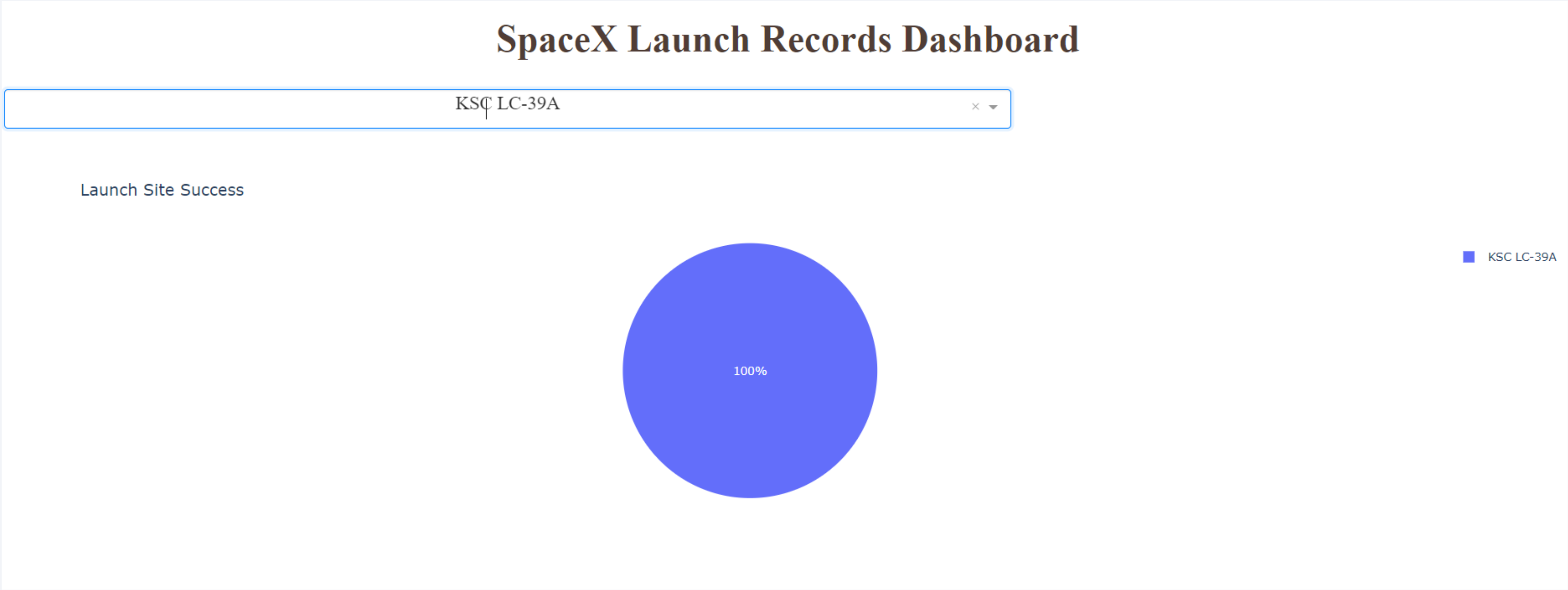
Build a Dashboard with Plotly Dash

Launch Sites Success Ratio

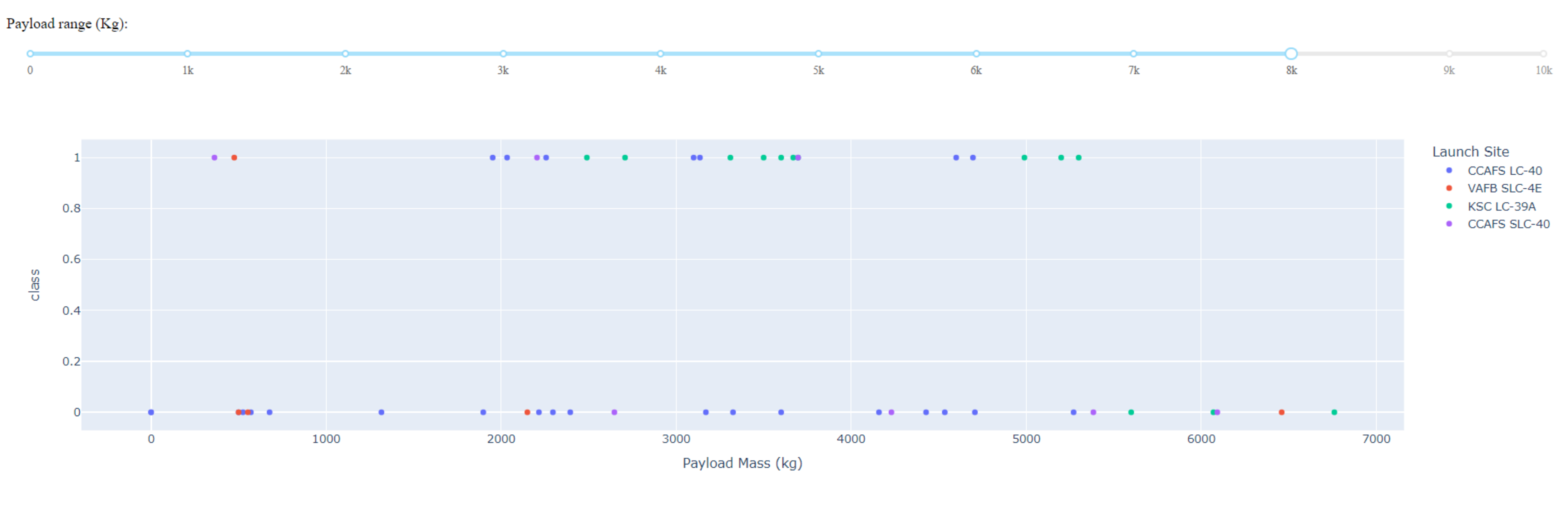


Site KSC-LC-39A has the highest success percentage

Launch Site with Highest Success Ratio



Payload vs Launch Outcome

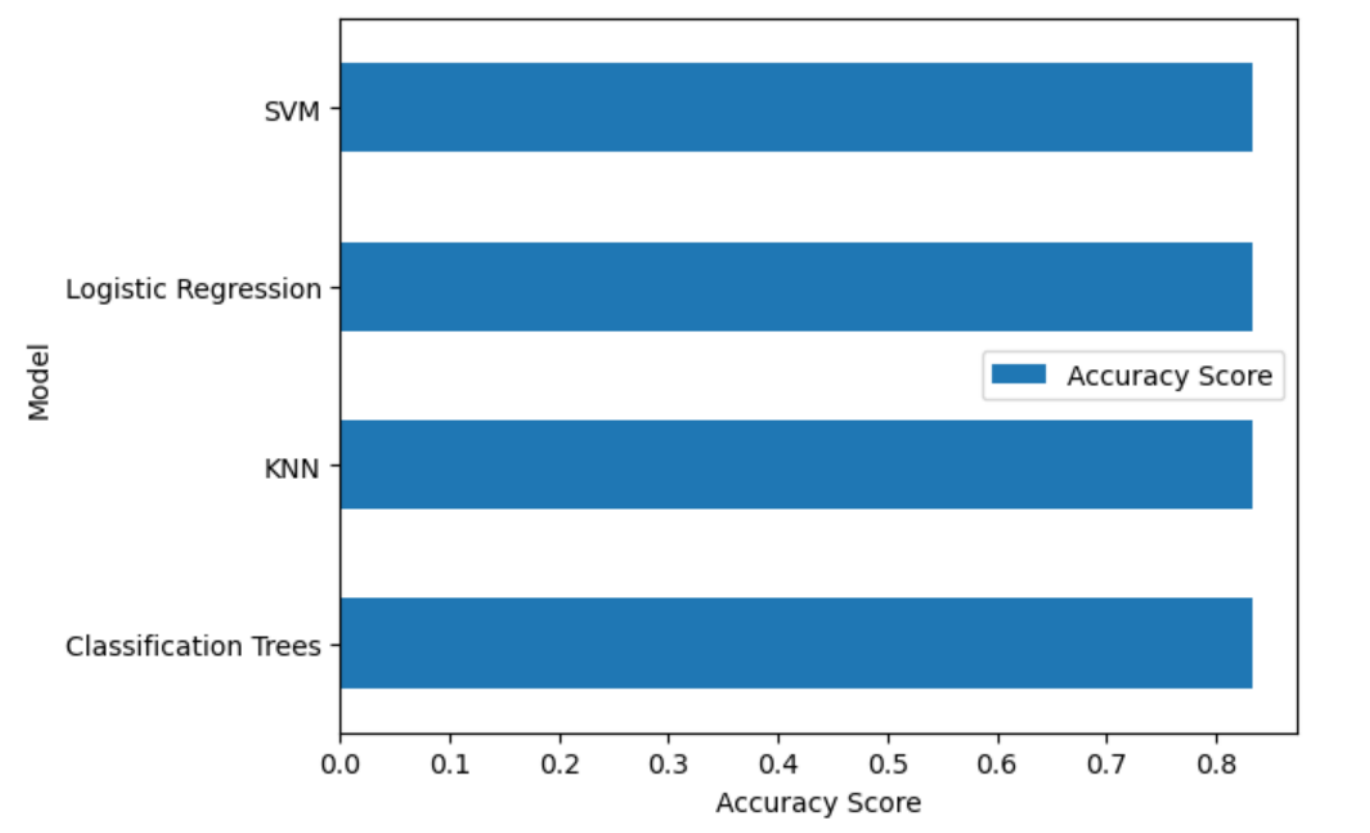


- Payload range between 2000 and 4000 have the highest successful landings

Section 5

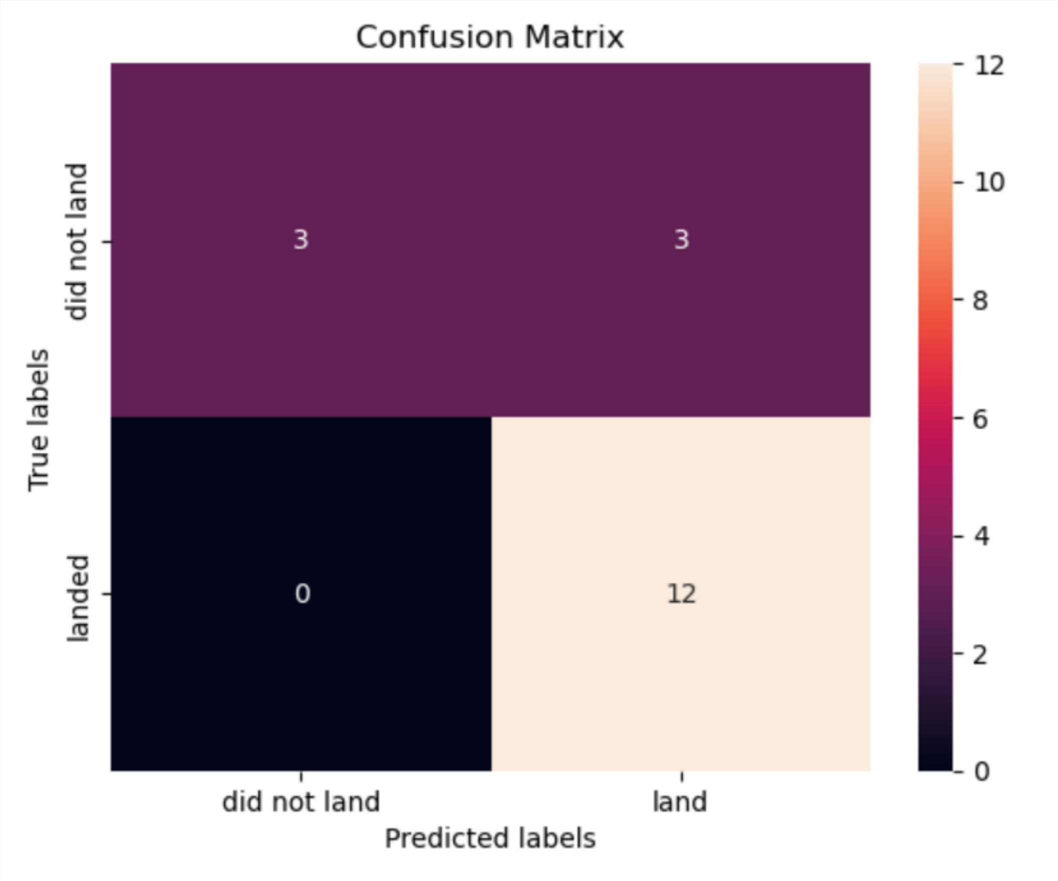
Predictive Analysis (Classification)

Classification Accuracy



Each model has an equal accuracy

Confusion Matrix



- From the confusion matrix we can see that our model is quite accuracy. Out of 15 predictions for a successful landing, 12 times, the model has predicted it successfully and out of 3 failed landings, we have predicted all of them correctly.

Conclusions

- Payload, Launch site and orbit are among the main predictors.
- Payload range between 2000 and 4000 have the highest successful landings
- Site KSC-LC-39A has the highest success percentage
- SSO, HEO, GEO and ES-L1 orbits have a 100% success rate
- For all launch sites, heavy payloads have a greater success rate for all launch sites
- Each model has an equal accuracy
- We can successfully predict the outcome at a percentage of 83%

Appendix

- https://github.com/hasnainnaeem93/IBMFinalProject

Thank you!