Winning Space Race with Data Science

Mahimai Raja J
30.12.2022

Outline

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

Executive Summary

- The data is collecting using web scrapping technique in python. Then the missing data is filled appropriately using the central tendency measures. After data is cleaning perfectly the data is visualized using folium library. The the hyperparameters were computed using Grid Search CV. Then multiple machine learning models were comparatively evaluated based on the performance.
- Using the real time dashboard by plotly dash we can clearly understand the story of the data. The interactive dashboard enables the user to understand the data in visual manner. Using multiple models on our data we can understand its nature. And finally we can come up with logistic regression as the best model of now with 84% training accuracy and 83%testing accuracy.

Introduction

- The project is to make predict whether the spacecraft will successfully launch or not. So that we are making our own data from internet using web scrapping with the help of beautiful soup in python. Then we prepared and processed the data. Then we compared by building multiple models using the graph and measured the performance.
- So using best model we are going to predict whether the spacecraft with the given environment condition will launch successfully or not.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - The data is collected using beautiful-soup in python.
- Perform data wrangling
  - The missing data is imputed using appropriate central tendency measures. Categorical variables are transformed using one-hot encoding. Then all the values were standardized using min-max scalar method.
- Perform exploratory data analysis (EDA) using visualization and SQL
  - The data from the database db2 is analyzed using magic sql on jupyter notebook.

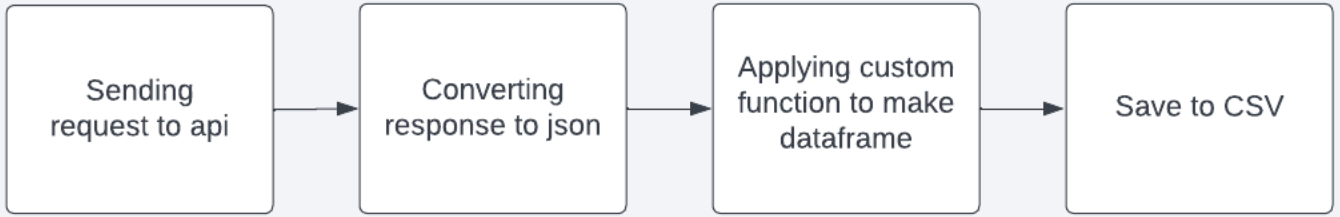
- Perform interactive visual analytics using Folium and Plotly Dash
    - The data is visualized using plotly dash a interactive dashboard in python.
- Perform predictive analysis using classification models
    - Based on the data selectively four model were test on the data and analysed with its performance over the testing and training data.

Data Collection

- Data collection is the process of gathering and measuring information on target variables in a established system, which then enables one to answer relevant questions and evaluate o
- In this project I have two methods to collect the spacex data by web scrapping and using the spacex api.
- I have attached the methodology and flow chart of the data collection process below.

Data Collection – SpaceX API

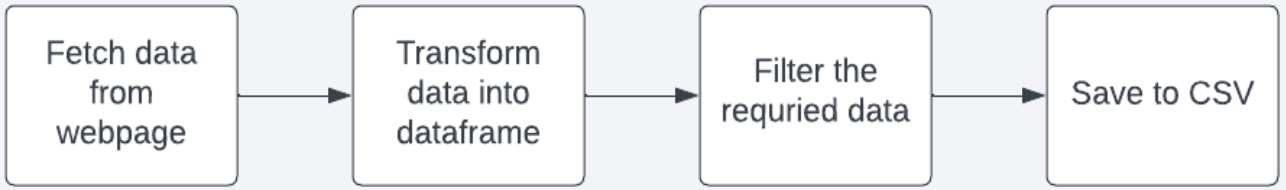
- Send request to the spacex api and store and response in json file. Then process the json with custom functions to make a dataframe and store it in csv.



- Github Link : https://github.com/mahimai-raja/project-spacex/blob/main/Week-01/notebookOne.ipynb

Data Collection - Scraping

- Data from the Wikipedia page is fetched using beautifulsoup and then converted into data frame and stored in the CSV.



- Github Link : https://github.com/mahimai-raja/project-spacex/blob/main/Week-01/notebookOne.ipynb

Data Wrangling

- The missing data is filled appropriately using central tendency methods. The categorical values are converted into numerical values using one-hot encoding. Then all the records were standardized using min-max scalar method.
- Github link : https://github.com/mahimai-raja/project-spacex/blob/main/Week-02/exploratoryDA.ipynb

EDA with Data Visualization

- I have scatter plot to find the correlation of data. And used barchart to visualize the categorical variables. I have used line plot to visualize the time defined features.
- Github url : https://github.com/mahimai-raja/project-spacex/blob/main/Week-02/dataVisualization.ipynb

EDA with SQL

- Used Select query to retrieve the data.
- Used Where clause to filter the required data.
- Used string operators such as ‘%’ and ‘_’
- Github Link : https://github.com/mahimai-raja/project-spacex/blob/main/Week-02/exploratoryDA.ipynb

Build an Interactive Map with Folium

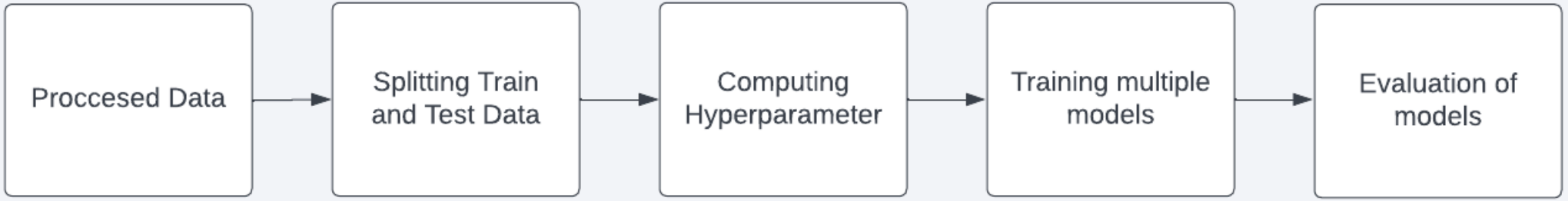
- Markers – Used to identify the location easily.
- Circle – Used to cluster a group of markers.
- Line – Used to measure distance visually in a map.
- The visual representation of the location using folium map is used to understand the spread of launch station and its significance.
- Github Link : https://github.com/mahimai-raja/project-spacex/blob/main/Week-03/site_location_folium_maps.ipynb

Build a Dashboard with Plotly Dash

- I have used pie chart and scatter plot to visualize the given data interactively using dash library in python.
- The pie chart is used to represent the proportional ration of the sample data. And scatter plot helps to identify the correlation between the attributes.
- Github Link : https://github.com/mahimai-raja/project-spacex/tree/main/Week-03/plotlyDashboard

Predictive Analysis (Classification)

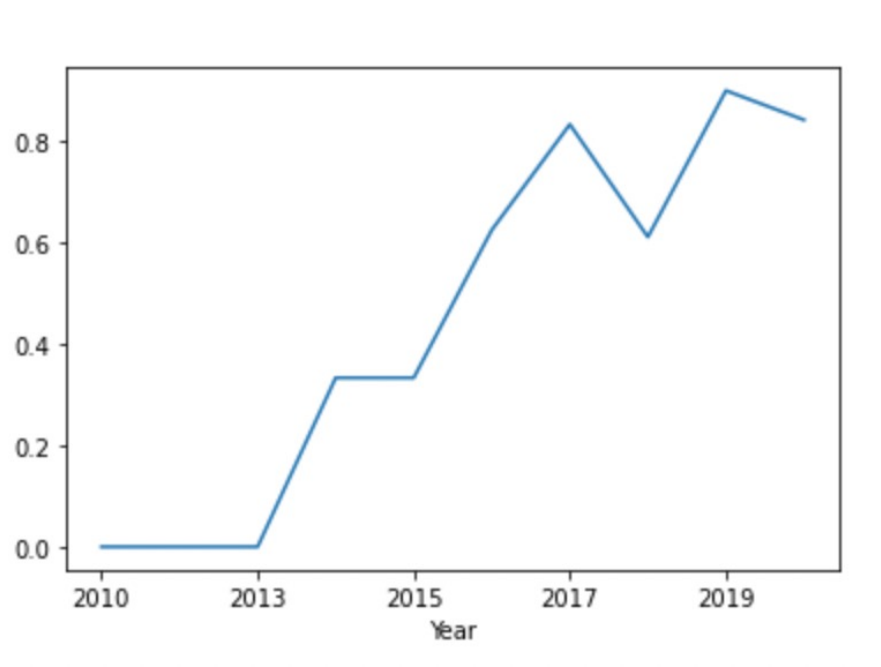
- The processing data is splitted into training and testing data. Then using grid search cv hyperparameters were computed. The multiple models were trained on the same data and evaluated using classification metrics.
- Github Link : https://github.com/mahimai-raja/project-spacex/blob/main/Week-04/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb



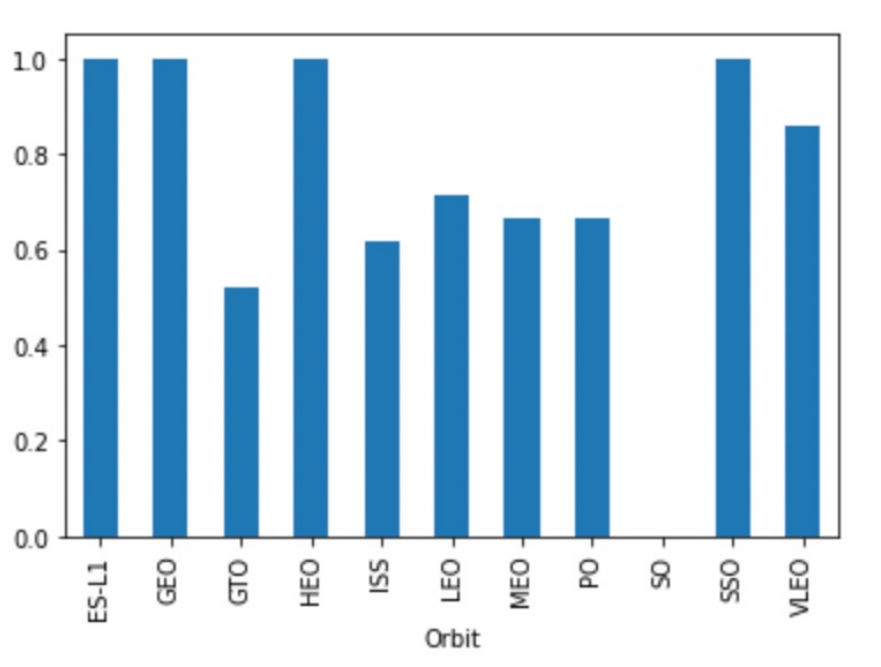
Results

- Exploratory data analysis results

Line Plot :



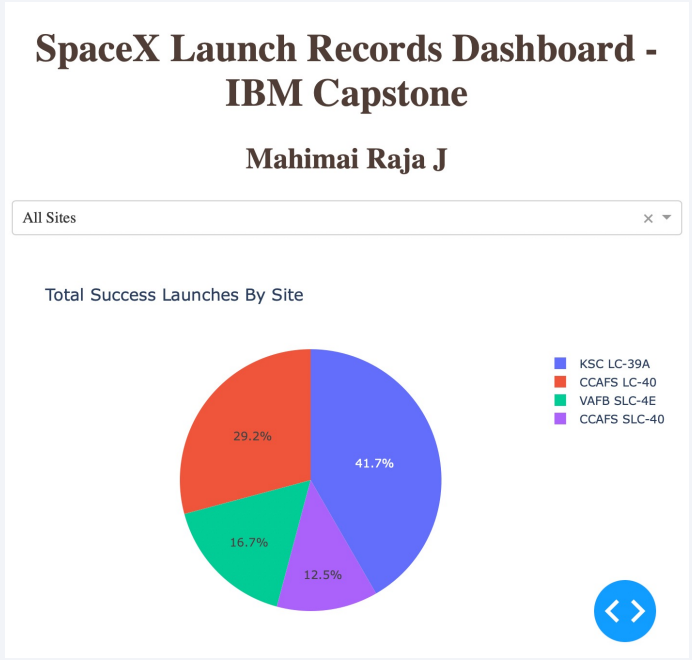
Bar Chart :



Results

- Interactive analytics demo in screenshots

Pie Chart :



Scatter Plot

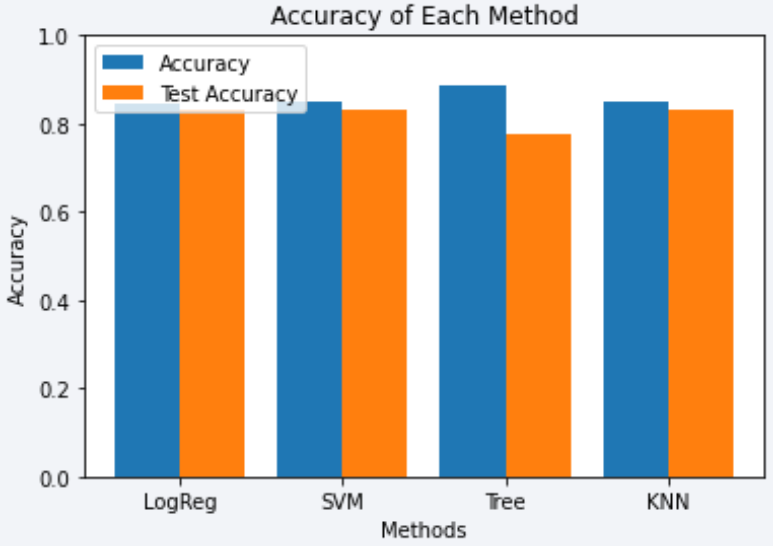
Results

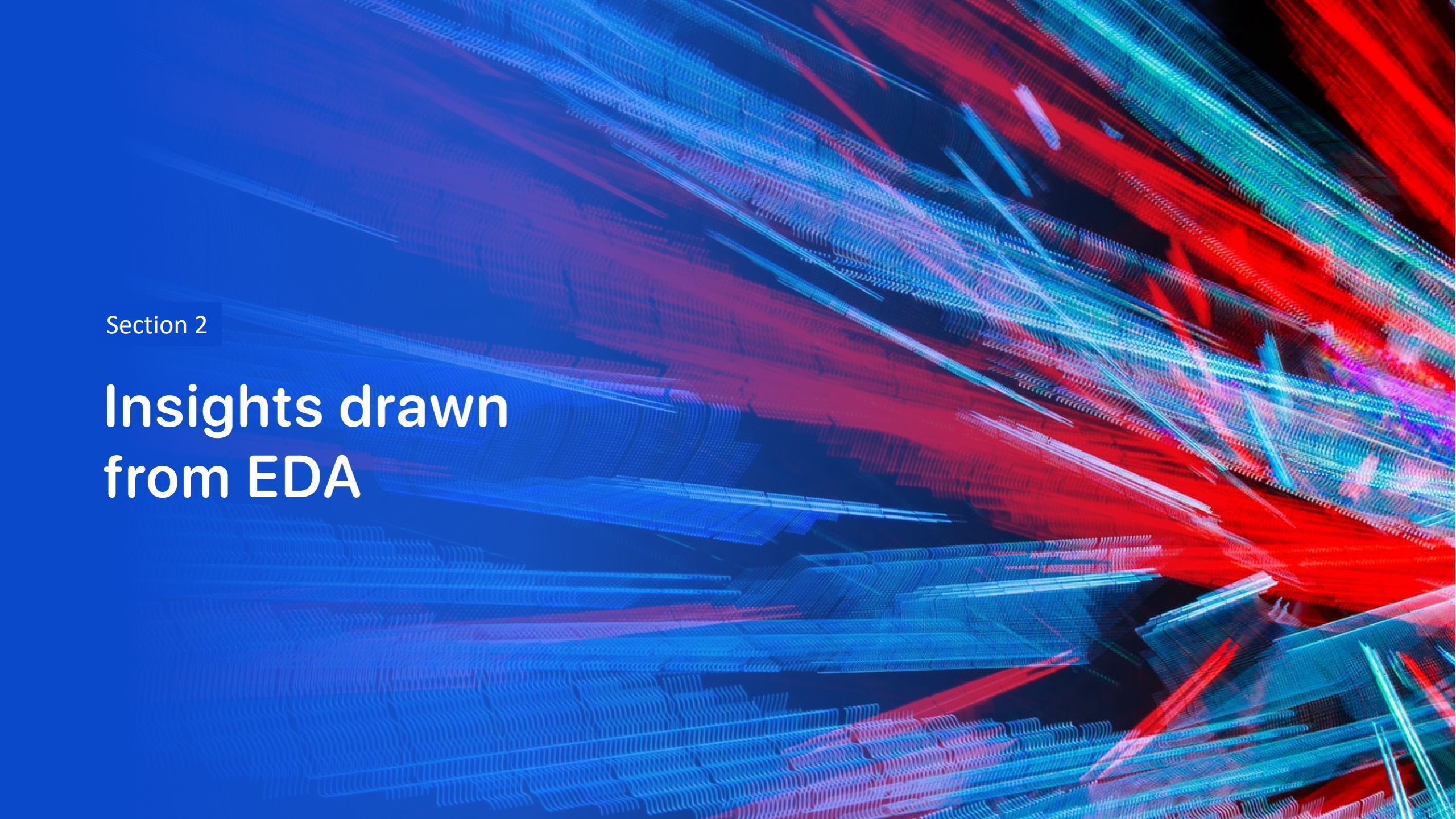
- Predictive analysis results

Comparative Analysis

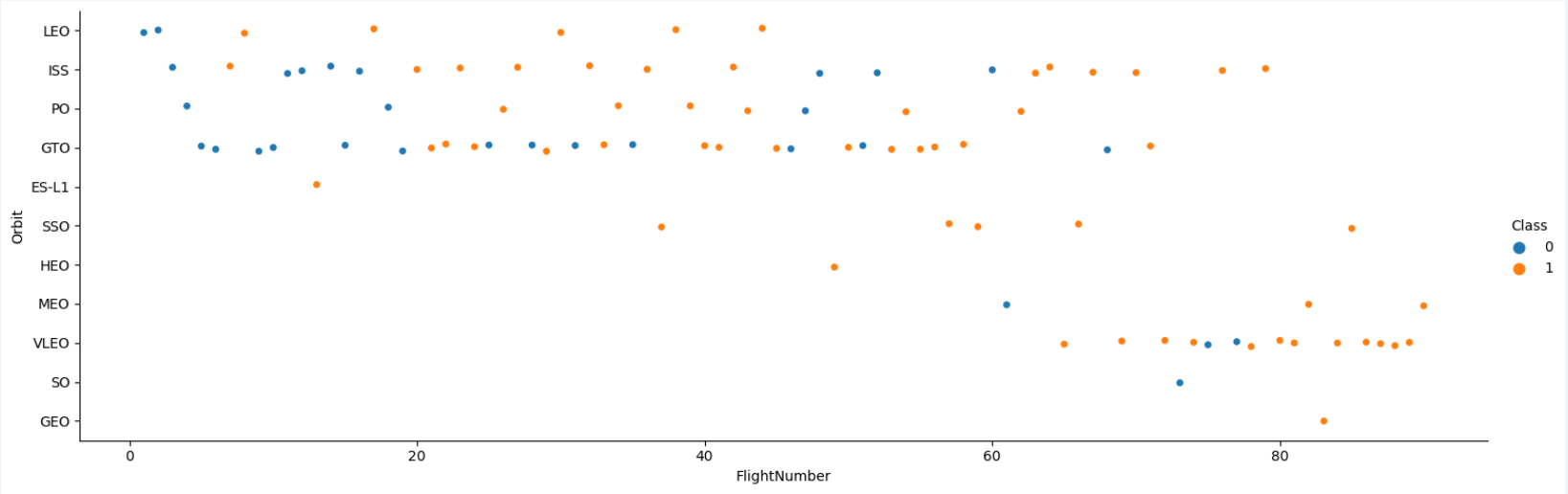
Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.88571	0.77778
KNN	0.84821	0.83333

Comparison Graph



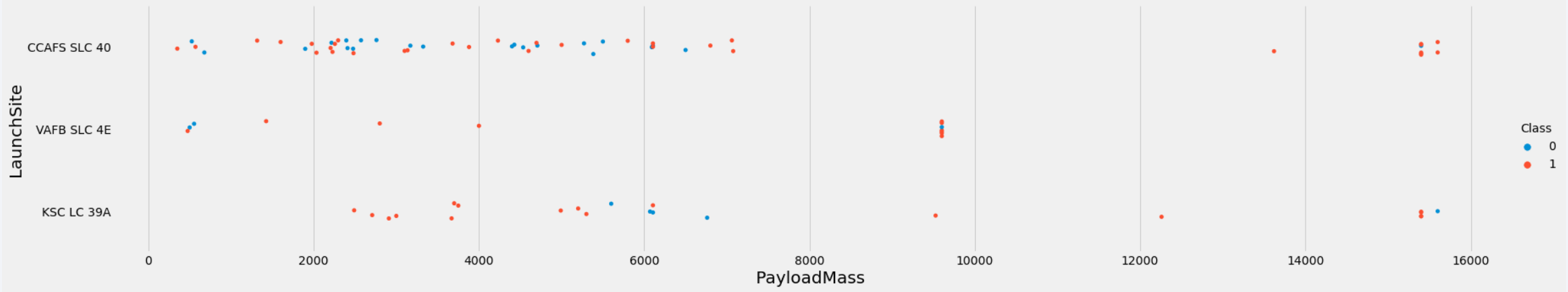

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



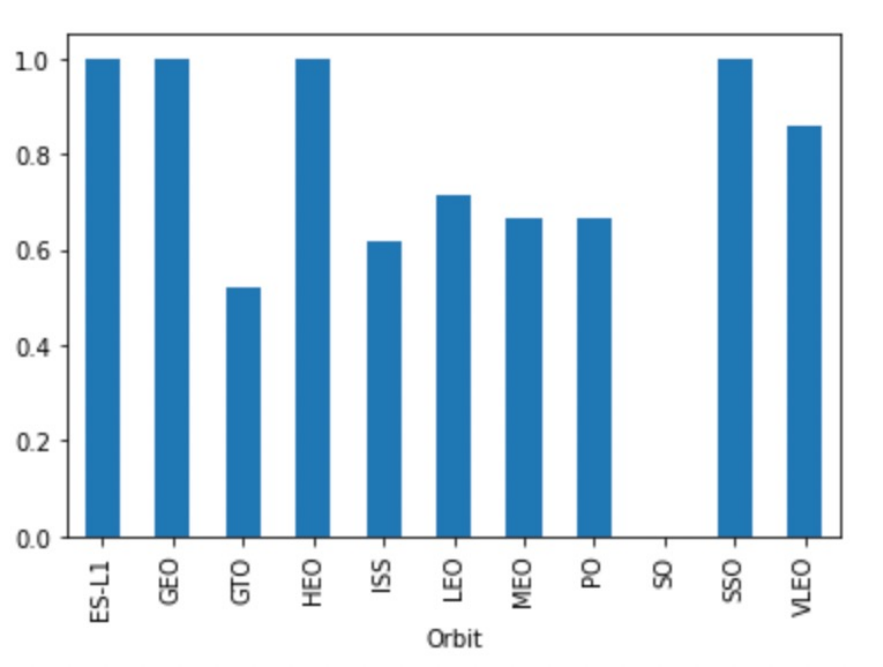
- From the above graph we can understand that VLEO launch side has high success rate comparatively.

Payload vs. Launch Site



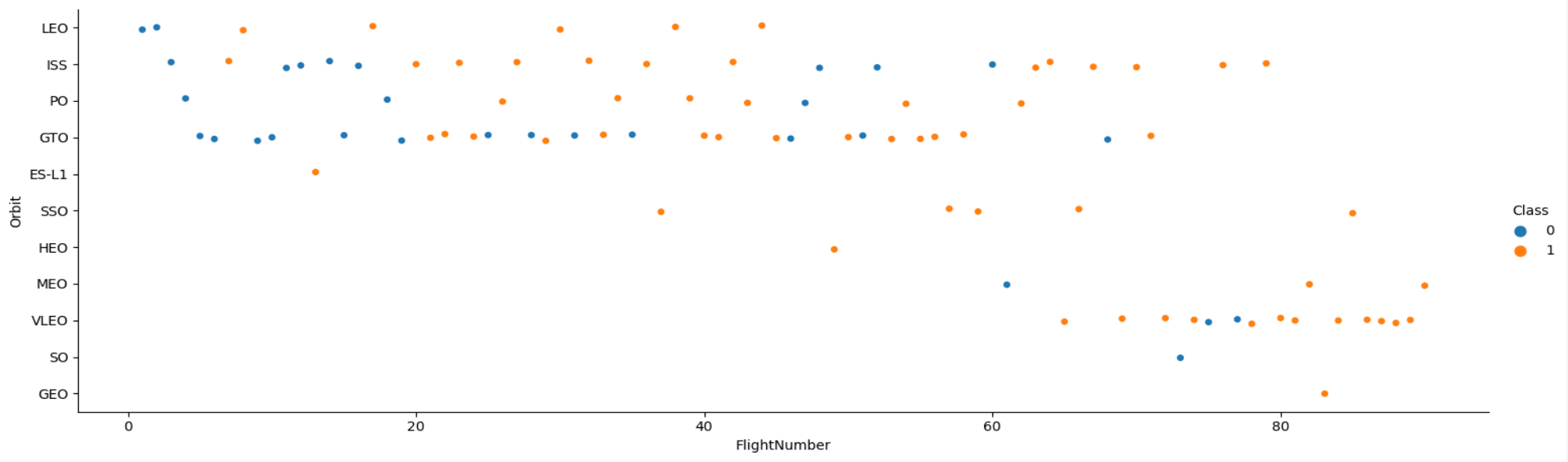
- From the scatter plot we can understand most of the launch sites were used to launch low weight payloads and the payloads above 10,000 mass have a high success rate.

Success Rate vs. Orbit Type



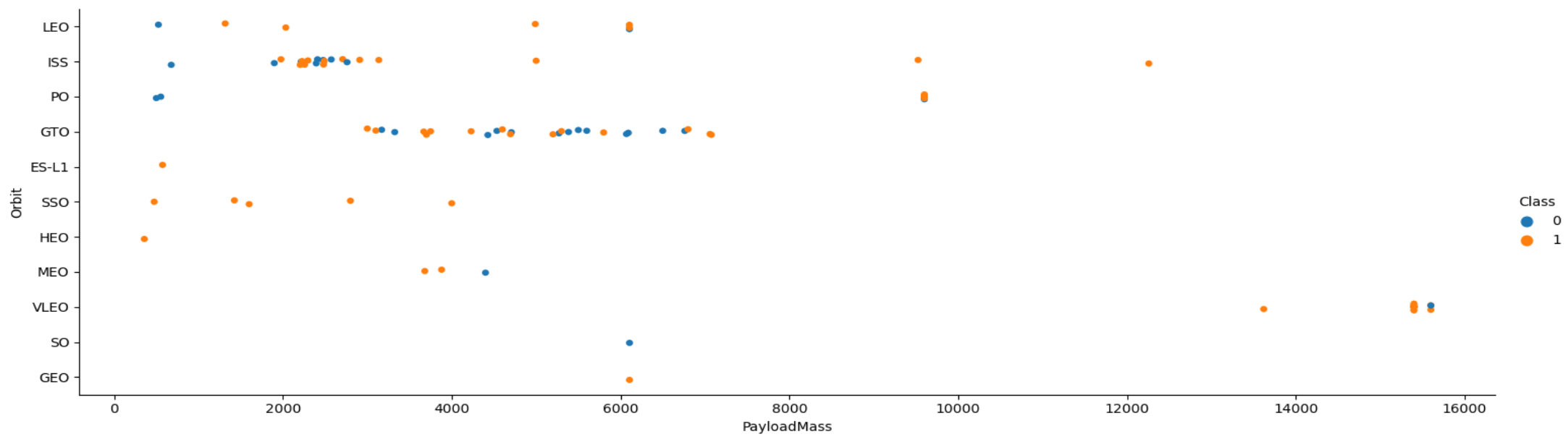
- From the above bar chart we can understand that the launch site ‘SO’ have zero success rate.

Flight Number vs. Orbit Type



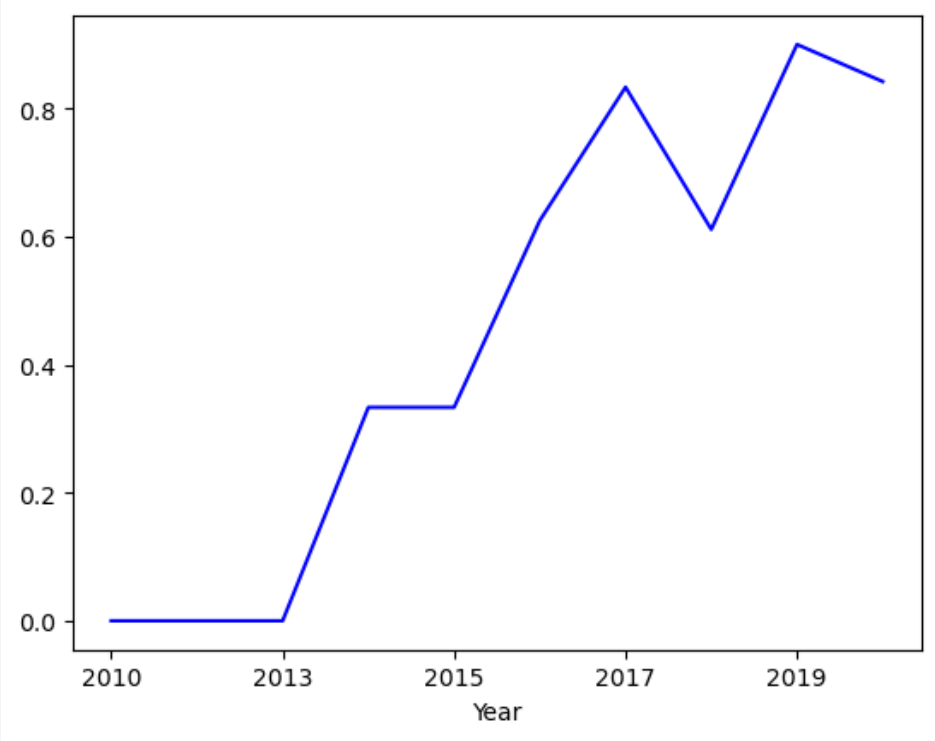
- From the above scatter plot we can understand that as the flight number increases the success rate is also more or less comparatively higher.

Payload vs. Orbit Type



- From the above graph we can understand that most of the payload were of lower mass comparatively.

Launch Success Yearly Trend



- From the above graph we can understand that the success rate is gradually increasing every year.

All Launch Site Names

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

- sql SELECT DISTINCT launch_site from space01;
- We can understand that there are four unique launch sites in the given dataset.

Launch Site Names Begin with 'CCA'

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

- sql SELECT * from space01 where launch_site LIKE 'CCA%' limit 5;
- We can undertand that the launch site starts with cca are usually weight less.

Total Payload Mass

payload_mass
45596

- `sql select sum(payload_mass__kg_) as Payload_Mass from SPACE01 where customer LIKE 'NASA (CRS)';`
- The sum of all the payload mass is 45,596

Average Payload Mass by F9 v1.1

Average Mass
2928

- `sql select avg(payload_mass__kg_) from SPACE01 WHERE BOOSTER_VERSION = 'F9 v1.1';`
- The average mass of all payload is 2928.

First Successful Ground Landing Date

First Success
01-05-2017

- `sql select min(DATE) from SPACE01 where landing__outcome LIKE 'Success (ground pad)';`
- Only from 2017 the ground pad launches were success

Successful Drone Ship Landing with Payload between 4000 and 6000

booster_version

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

- `sql select booster_version from SPACE01 where landing__outcome LIKE 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000`
- There were only four booster versions between 4000 to 6000 mass.

Total Number of Successful and Failure Mission Outcomes

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

- `sql SELECT MISSION_OUTCOME, COUNT(*) AS Total_Number FROM SPACE01 GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;`
- The number of success record is higher in the given dataset.

Boosters Carried Maximum Payload

booster_version
F9 B5B1048.4
F9 B5B1049.4
F9 B5B1051.3
F9 B5B1056.4
F9 B5B1048.5
F9 B5B1051.4
F9 B5B1049.5
F9 B5B1060.2
F9 B5B1058.3
F9 B5B1051.6
F9 B5B1060.3
F9 B5 B1049.7

- `sql select booster_version from SPACE01 where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACE01);`
- There were12 booster version with maximum weight.

2015 Launch Records

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

- `sql select booster_version, launch_site from SPACE01 WHERE LANDING__OUTCOME = 'Failure (drone ship)' and DATE LIKE '%2015';`
- In2015 there were two booster version led to failure by drone ship.

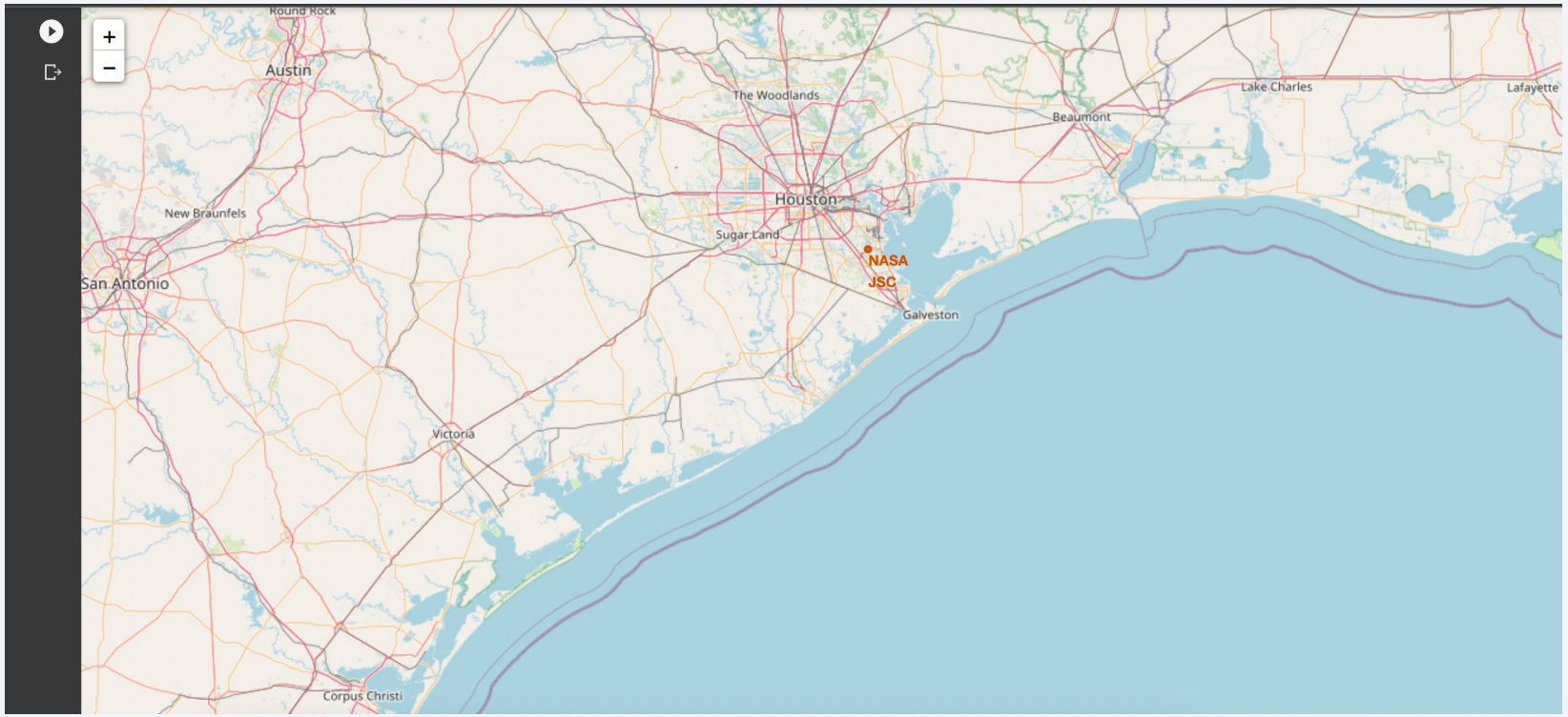
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

landing__outcome	Total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

- `sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACE01 WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME order by LANDING__OUTCOME;`
- No attempt is higher between 2004 and 2020.

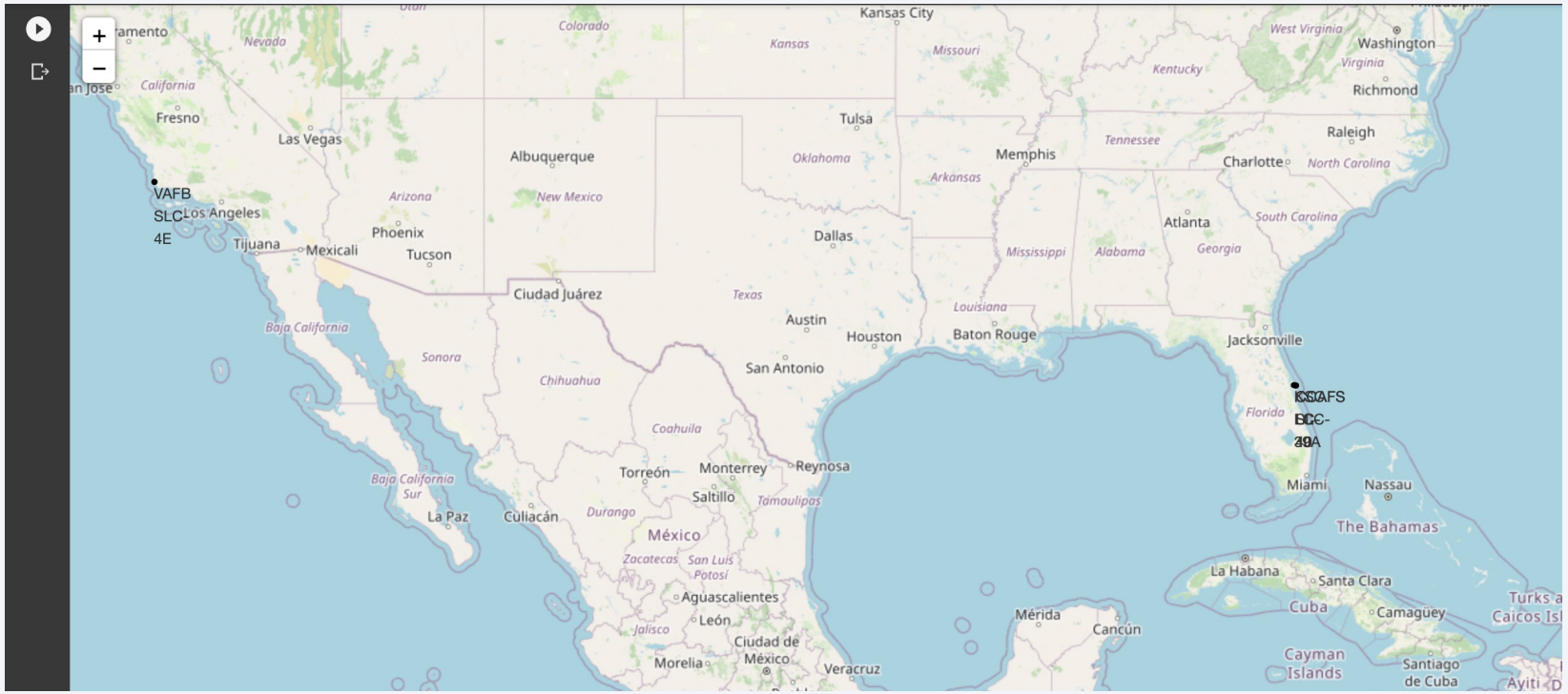

Section 3
Launch Sites
Proximities Analysis

NASA Johnson Space Center



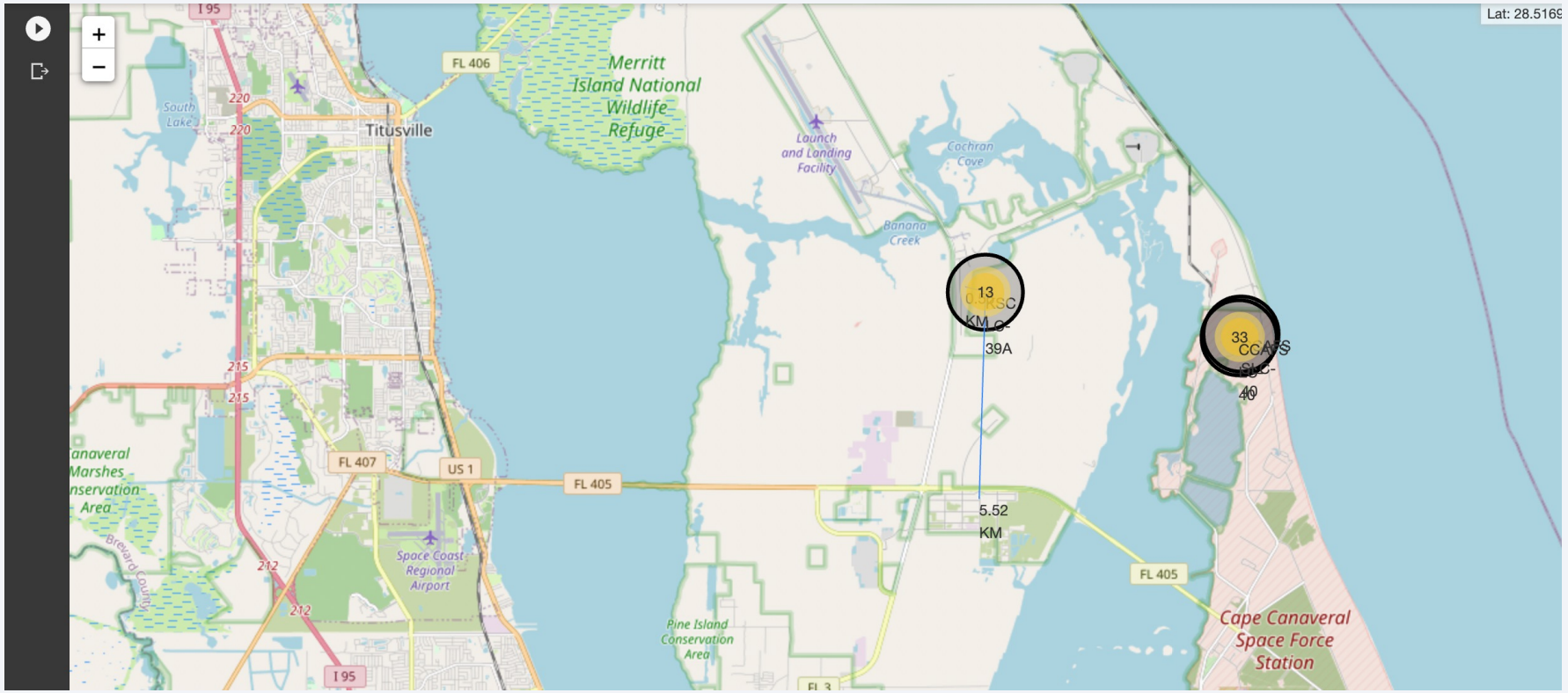
- Here we used circle and marked to specify the NASA Johnson Space Center.

Launch Sites

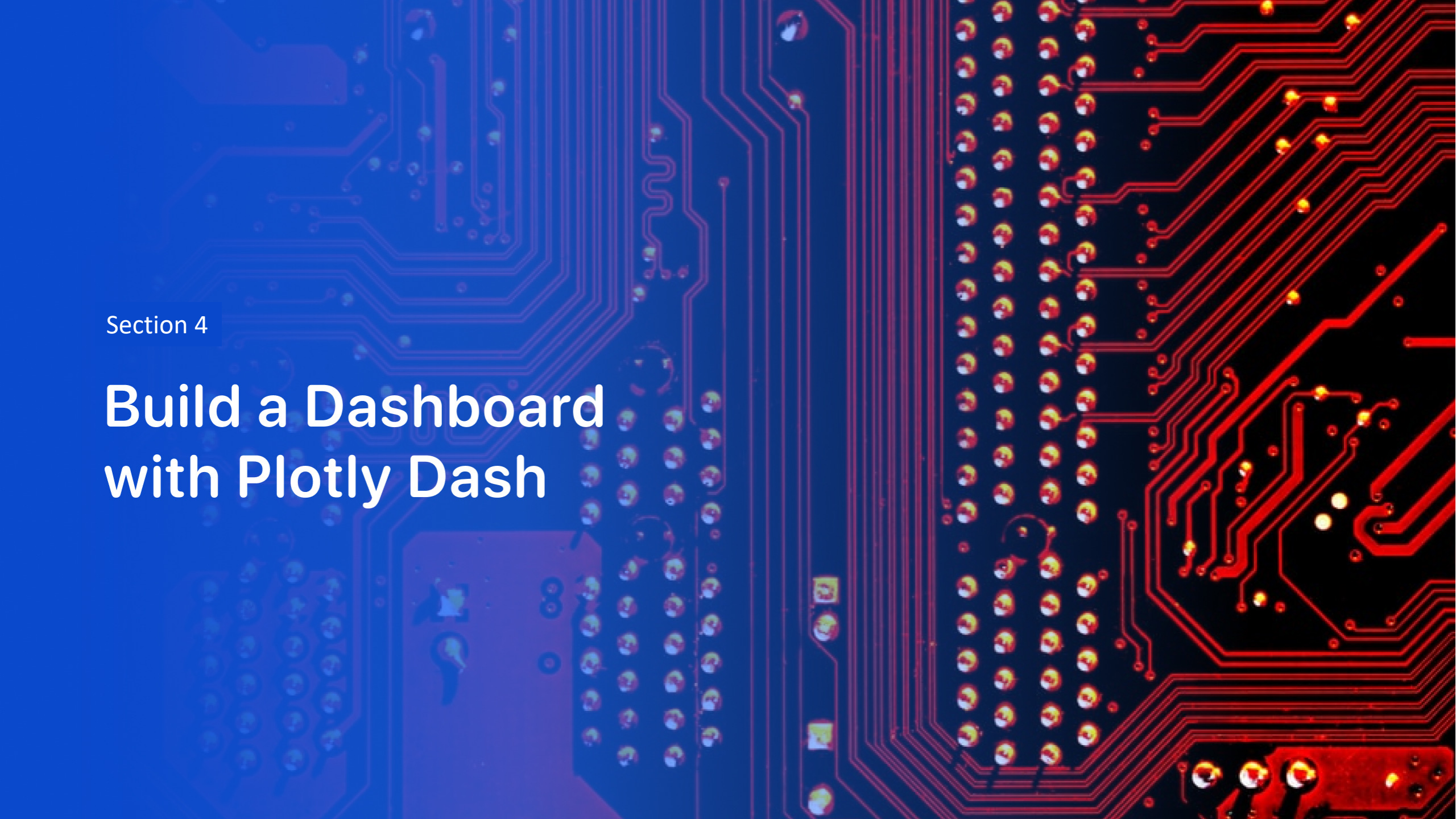


- Here we have used circle and marker to specify all the unique launch sites.

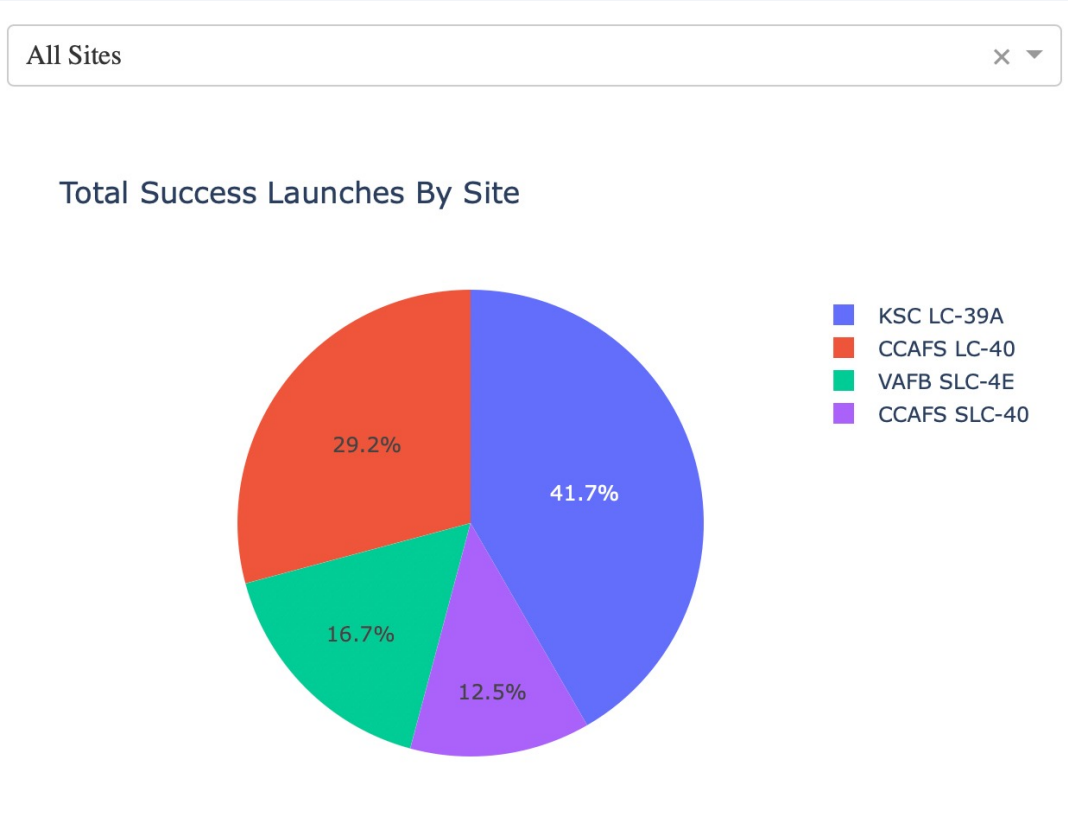
Distance to closest city



- Here we have clusters the nearby launch sites and used distance marker to calculate the distance to nearest city, railway and highway

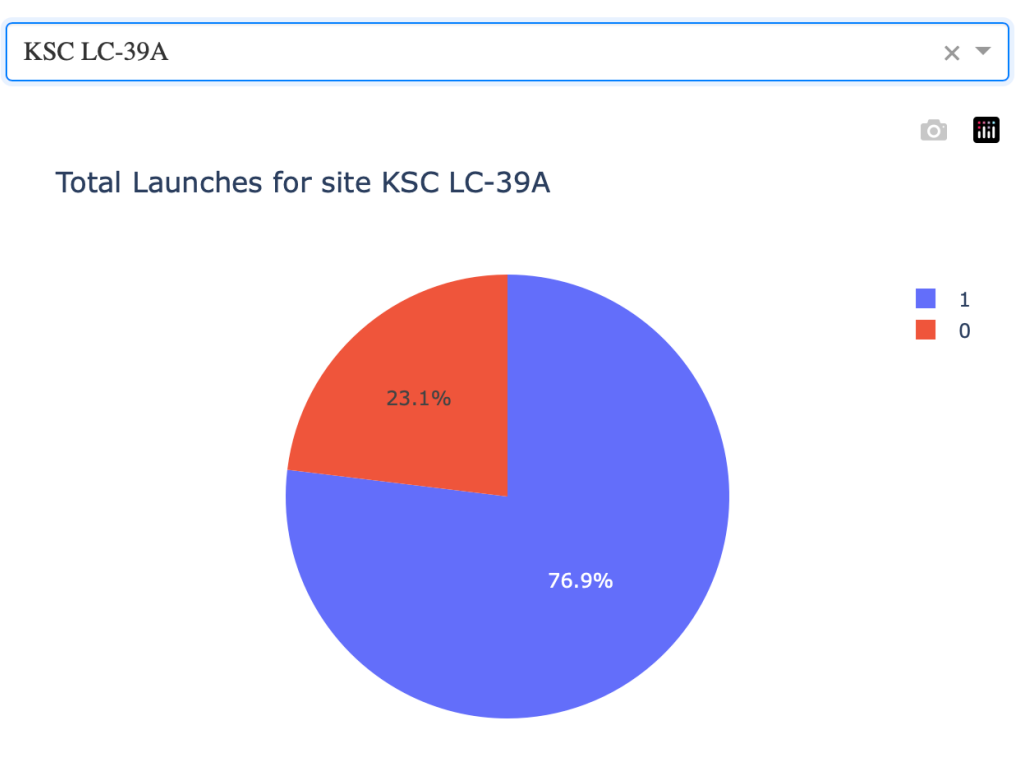

Section 4
Build a Dashboard with Plotly Dash

Success Rate – Pie Chart



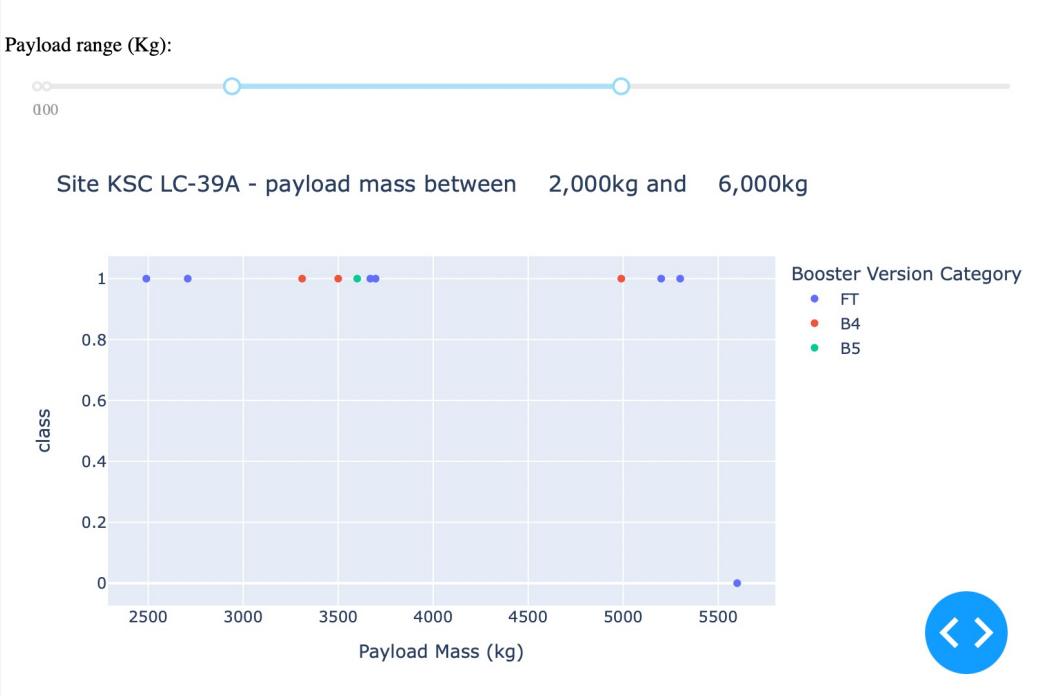
- From the above pie chart we can understand that KSC LC-39A has the higest success rate

Success Ration - KSC-LC-39A



- From the above pie chart we can understand that the failure rate is comparatively lesser.

Payload on Success – Scatter plot

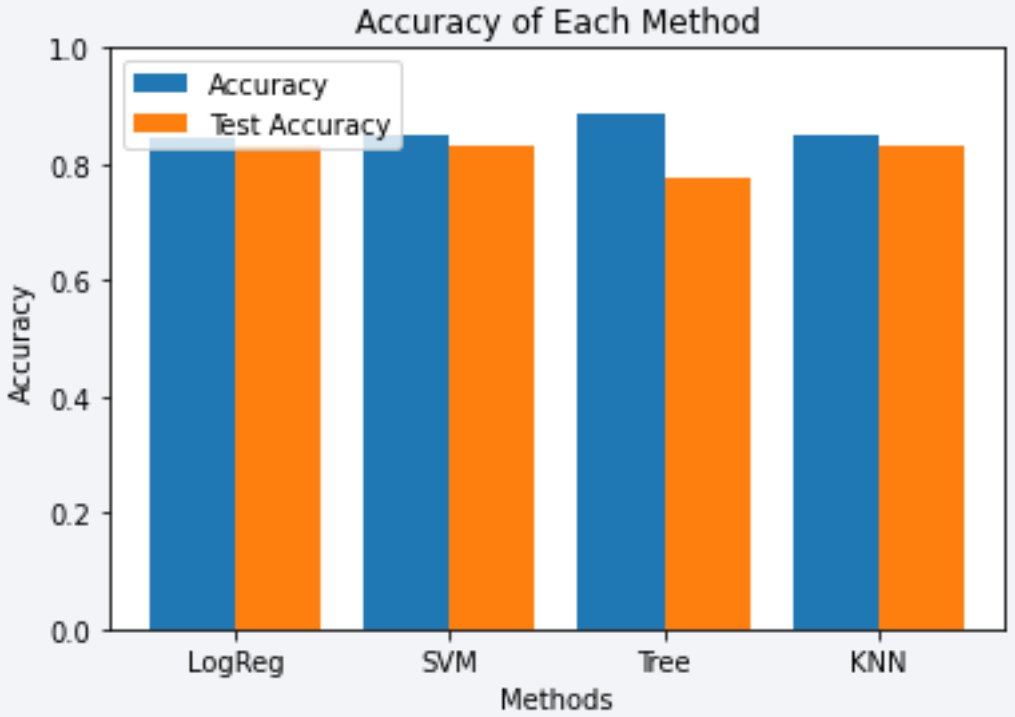


- From the above scatter plot we can understand that most of the successful payloads we lesser that 6000 unit mass.

Section 5

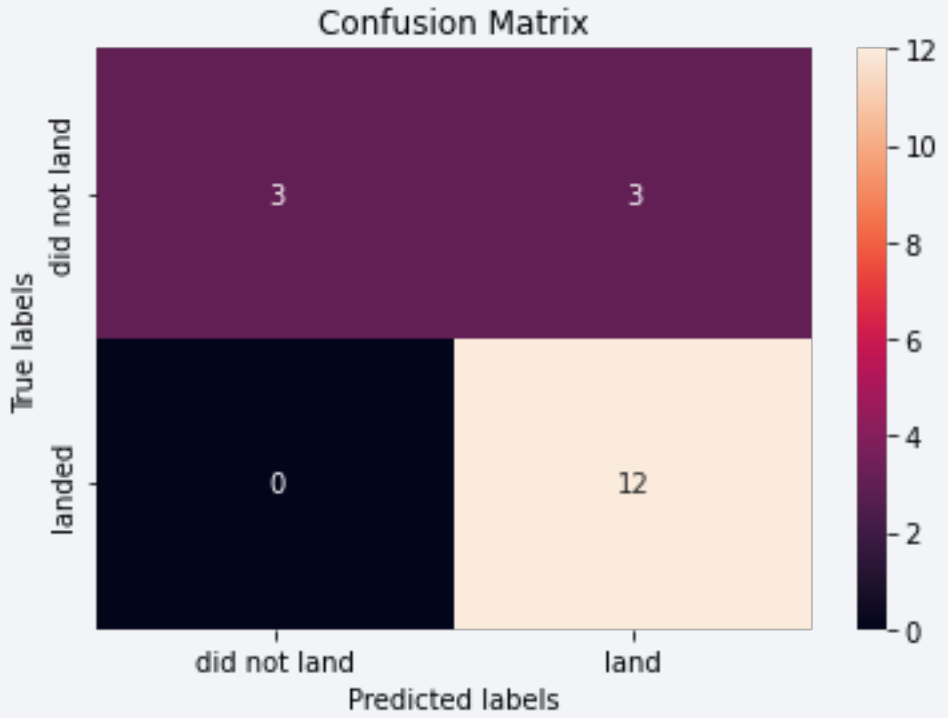
Predictive Analysis (Classification)

Classification Accuracy



- Here we have decision tree classifier with the highest accuracy but, It lacks in generalization. So we can go with the second best model logistic regression.

Confusion Matrix – Logistic Regression



- By looking into the confusion matrix we can understand that our logistic regression model has only few wrong predictions. And the overall performance is good.

Conclusions

- Orbit ES-L1, GEO, HEO, SSO has highest success rate.
- Success rates for spaceX launches has been increasing relatively with time and it looks like soon they will reach the required targer.
- Most of the successful payloads were lesser than 8000 unit mass.
- Logistic Regression is the best of the models that I have trained with the given dataset.

Appendix

- The model then converted into joblib file format and then hosted in ‘Pythion everywhere’ using flask framework.
- The live interactive dashboard is hosted using plotly dash and flask.

Thank you!