DATA SCIENCE CAPSTONE – WINING THE SPACE RACE

Lokesh B
19-Feb

Outline

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

Executive Summary

Summary of methodologies

- Data Collection using web scraping and SpaceX API.
- Data wrangline, Exploratory Data Analysis (EDA), data visualization
- Interactive visual analytics with folium
- Machine Learning Prediction using Logistic Regression, Decision Tree, SVM, KNN models.

Summary of all results

- All model produced similar results with accuracy rate of about 83.33%.
- All model overpredicted no of successful landings i.e. models predicted false positives.

Introduction

Project Background

- Spacex advertises Falcon 9 rocket launches on its website, with a cost of 62 million USD, with provides saving because other quote upwards of USD 165 Million, This is due to reusable stage one boosters. There fore it becomes important to determine that stage 1 boosters will land safely.

The objective

- is to evaluate the viability of the new company Space Y to compete with Space X.
- To compete with Space X, we need to estimate the cost of launches, which can be done by predicting successful landings of the first stage of rockets;
- This is done by using Machine Learning models on past Space X data to predict whether Stage 1 of rocket will be reused or not

METHODOLOGY

Methodology

Executive Summary

- **Data collection methodology:**
  - Data from Space X was obtained from 2 sources:
  - Space X API (https://api.spacexdata.com/v4/rockets/
  - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches)
- **Perform data wrangling**
  - Collected data was enriched by creating a landing outcome label based on outcome data after summarizing and analyzing features
- **Perform exploratory data analysis (EDA) using visualization and SQL**

Methodology

Executive Summary

- **Perform interactive visual analytics using Folium and Plotly Dash**
- **Perform predictive analysis using classification models**
    - Data collected was normalized, divided in training and test data sets and evaluated by four different classification models, being the accuracy of each model evaluated using different combinations of parameters which were selected using gridsearchcv.

Data Collection

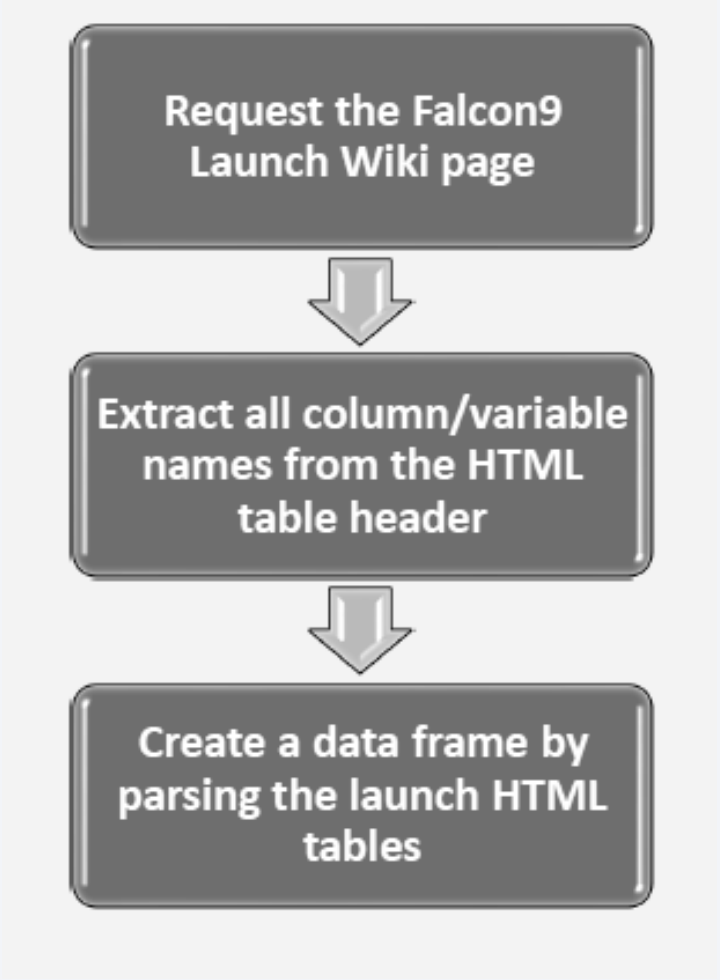
- Data from Space X was obtained from 2 sources:
- Space X API (https://api.spacexdata.com/v4/rockets/
- Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches)

Data Collection - SpaceX API

- SpaceX offers a public API from where data can be obtained and then used.
- This API was used according to the flowchart beside and then data is persisted.
- Source Code:- https://github.com/BhagwatiSharanKhare/Leaap/blob/main/Applied%20Data%20Science%20Capstone/1.%20Data%20Collection%20API.ipynb

Data Collection - Scraping

- Data from SpaceX launches can also be obtained from Wikipedia;
- Data are downloaded from Wikipedia according to the flowchart and then persisted.
- Source Code:- https://github.com/BhagwatiSharanKhare/Leaap/blob/main/Applied%20Data%20Science%20Capstone/2.%20Data%20Collection%20with%20Web%20Scraping.ipynb



Data Wrangling

•We Created a classification training label with landing outcomes where successful is 1 & failure is 0.

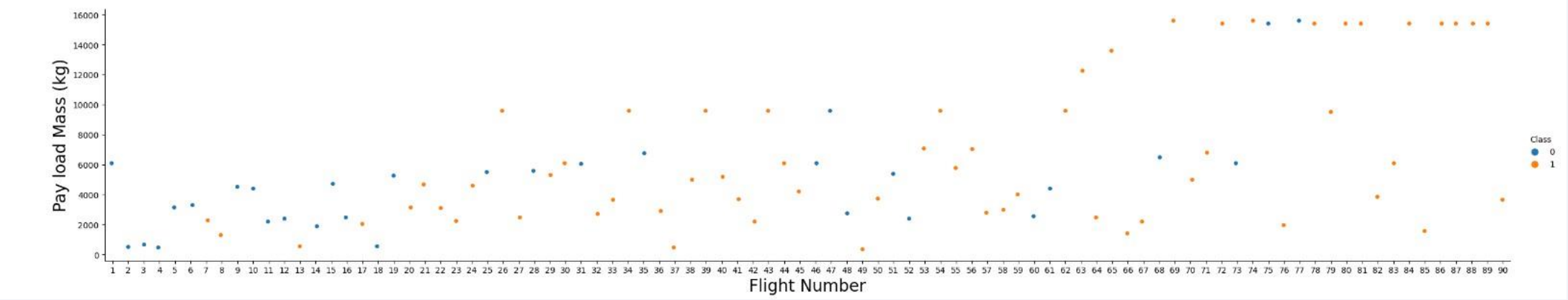
• If Outcome are True ASDS, True RTLS, & True Ocean class is set to 1.

•If Outcome are None None, False ASDS, None ASDS, False Ocean, False RTLS class is set to -> 0

•Source Code:-
https://github.com/BhagwatiSharanKhare/Leaap/blob/main/Applied%20Data%20Science%20Capstone/3.%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model
- Source Code:- https://github.com/BhagwatiSharanKhare/Leaap/blob/main/Applied%20Data%20Science%20Capstone/5.%20EDA%20with%20Visualization.ipynb



EDA with SQL

- Firstly the data was loaded in a IBM DB2 Database.

•Various Queries were performed using SQL magic function in jupyter notebook.

- Queries were made to get a better understanding of the dataset.

•Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

•Source Code:-
https://github.com/BhagwatiSharanKhare/Leaap/blob/main/Applied%20Data%20Science%20Capstone/4.%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

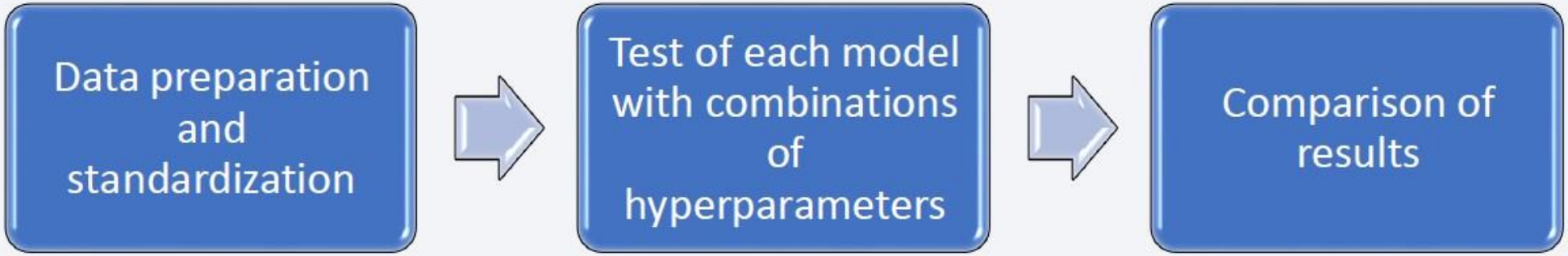
- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites;
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center;
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site;
  - Lines are used to indicate distances between two coordinates.
- Source Code:- https://github.com/BhagwatiSharanKhare/Leaap/blob/main/Applied%20Data%20Science%20Capstone/6.%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
- Percentage of successful launches by site as a pie chart
- Visualizing launches by site & Payload range as a scatter plot
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.
- Source Code:- https://github.com/BhagwatiSharanKhare/Leaap/blob/main/Applied%20Data%20Science%20Capstone/8.%20spacex_dash_app.py

Predictive Analysis (Classification)

- Four classification models were compared: logistic regression, support vector machine, decision tree and k nearest neighbors.
- Source Code:- https://github.com/BhagwatiSharanKhare/Leaap/blob/main/Applied%20Data%20Science%20Capstone/7.%20Machine%20Learning%20Prediction.ipynb



Results

- Exploratory data analysis results:
- Space X uses 4 different launch sites;
- The first launches were done to Space X itself and NASA;
- The average payload of F9 v1.1 booster is 2,928 kg;
- The first success landing outcome happened in 2015 fiver year after the first launch;
- Many Falcon 9 booster versions were successful at landing in drone ships having payload above the average;
- Almost 100% of mission outcomes were successful;
- Two booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015;
- The number of landing outcomes became as better as years passed.

Results

- Using interactive analytics was possible to identify that launch sites use to be in safety places, near sea, for example and have a good logistic infrastructure around.
- Most launches happens at east cost launch sites.
- Predictive Analysis showed that Decision Tree Classifier is the best model to predict successful landings, having accuracy over 87% and accuracy for test data over 83%

Section 2

Insights drawn from EDA

All Launch Site Names

- They are obtained by selecting distinct occurrences of “ launch_site ” values from the dataset.

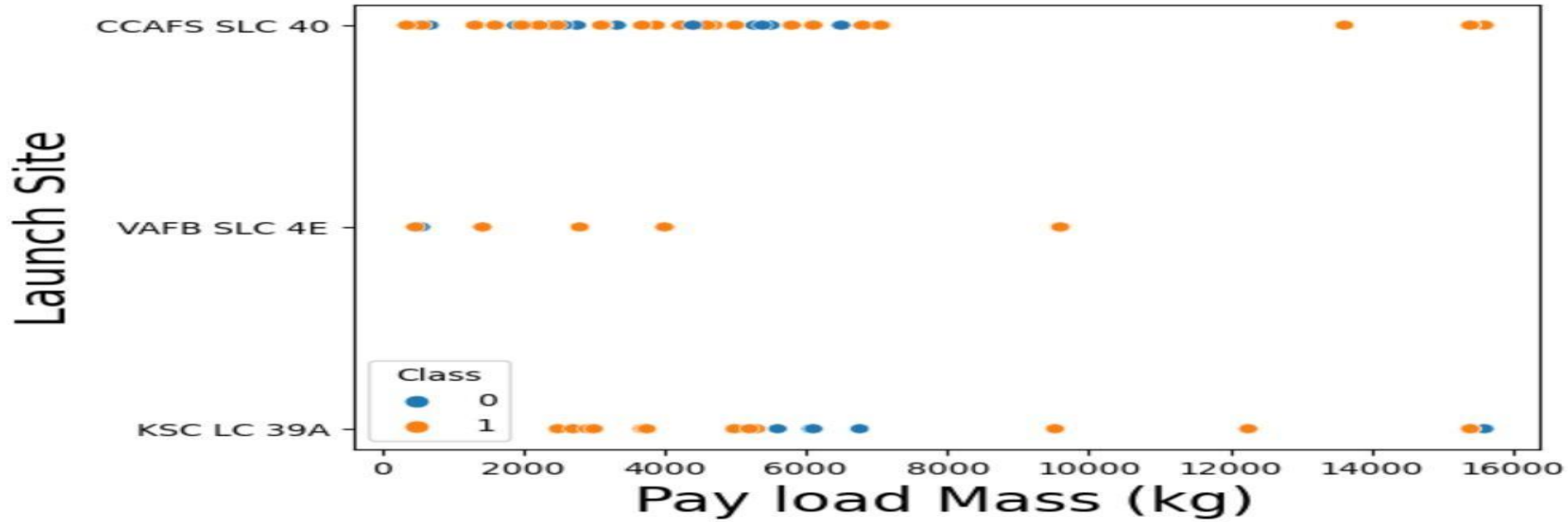
launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Total Payload Mass

- Total payload calculated above, by summing all payloads whose codes contain ‘ CRS’, which corresponds to NASA.

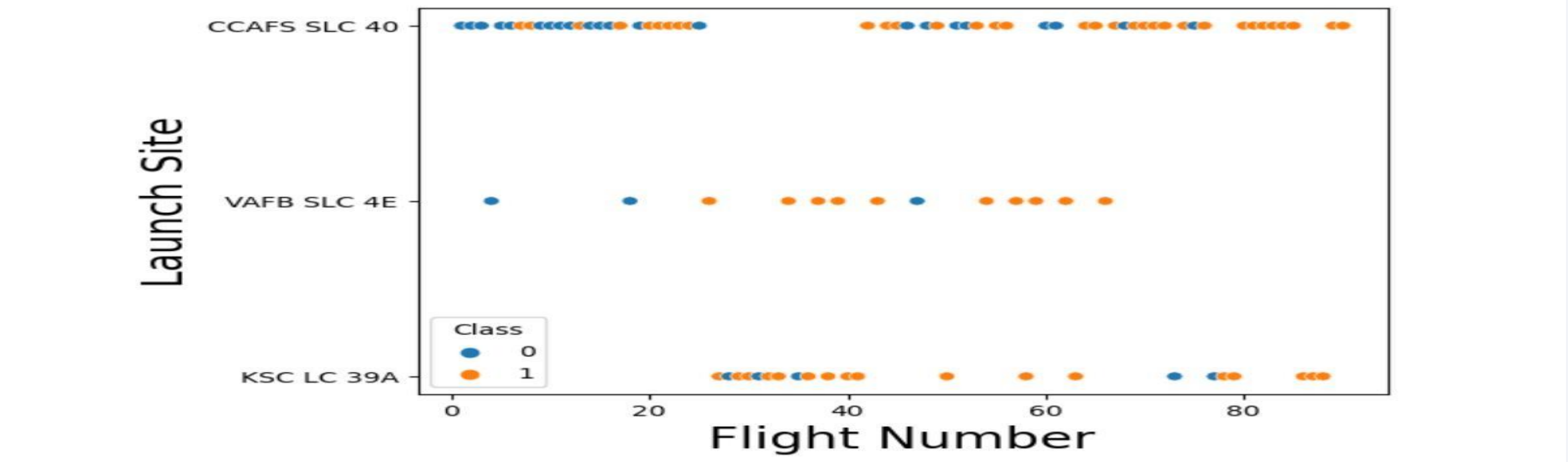
1
45596

Payload vs. Launch Site



- Payloads over 9,000kg (about the weight of a school bus) have excellent success rate;
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC39A launch sites.

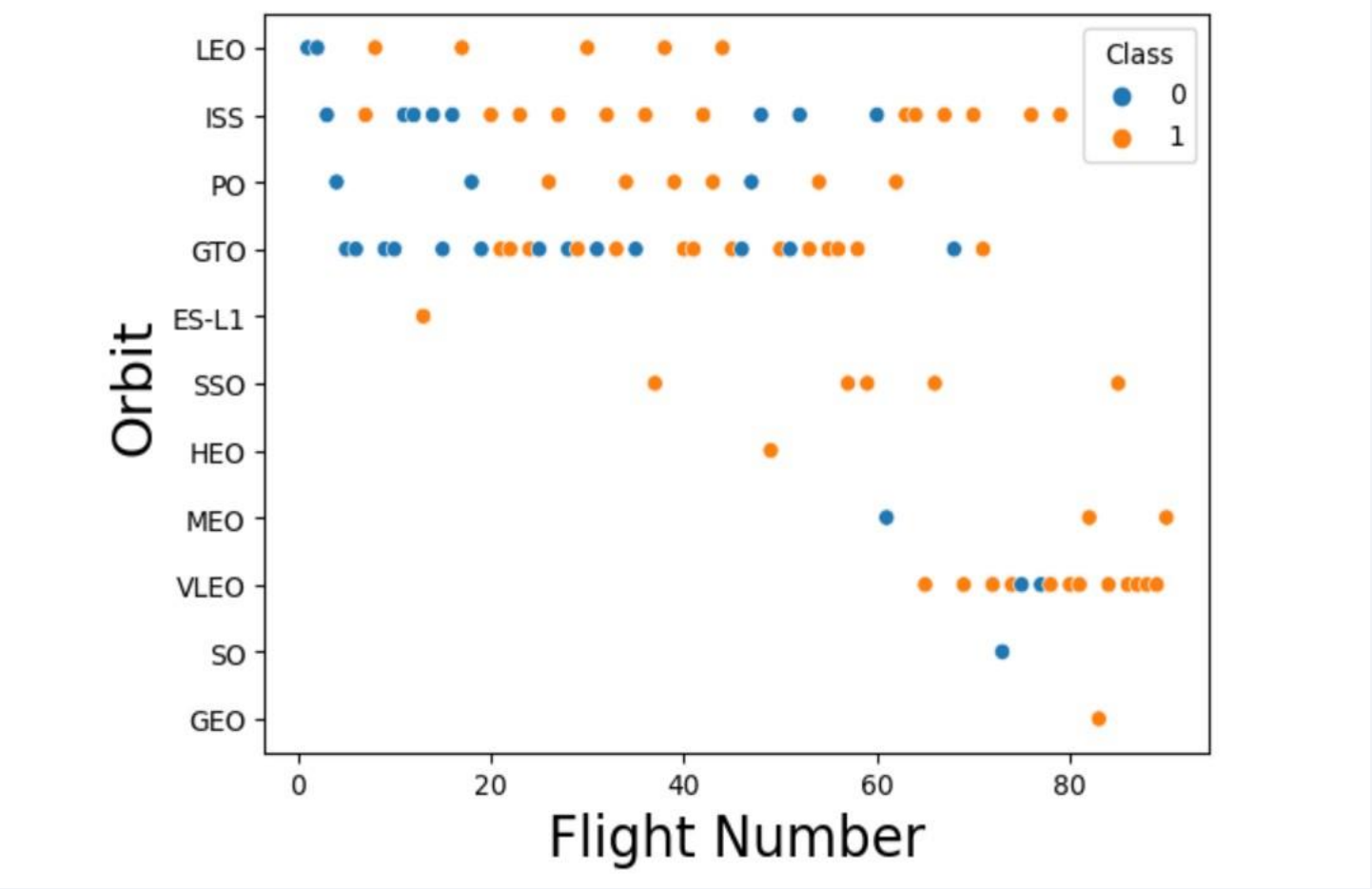
Flight Number vs. Launch Site



- According to the plot above, it's possible to verify that the best launch site nowadays is CCAF5 SLC 40, where most of recent launches were successful
- It's also possible to see that the general success rate improved over time.

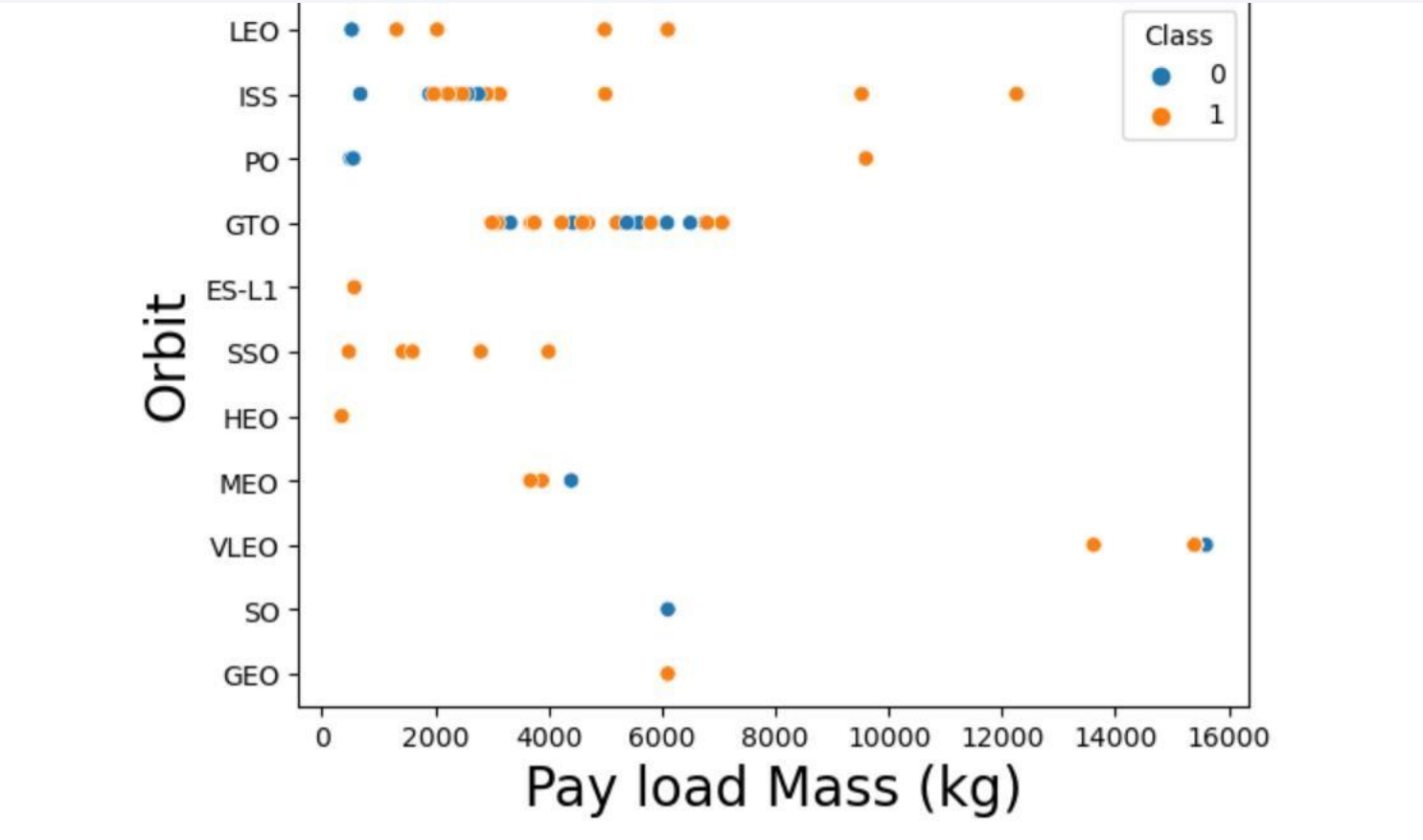
Flight Number vs. Orbit Type

- No of flights have increased recently for VLEO.
- Apparently, success rate improved over time to all orbits;



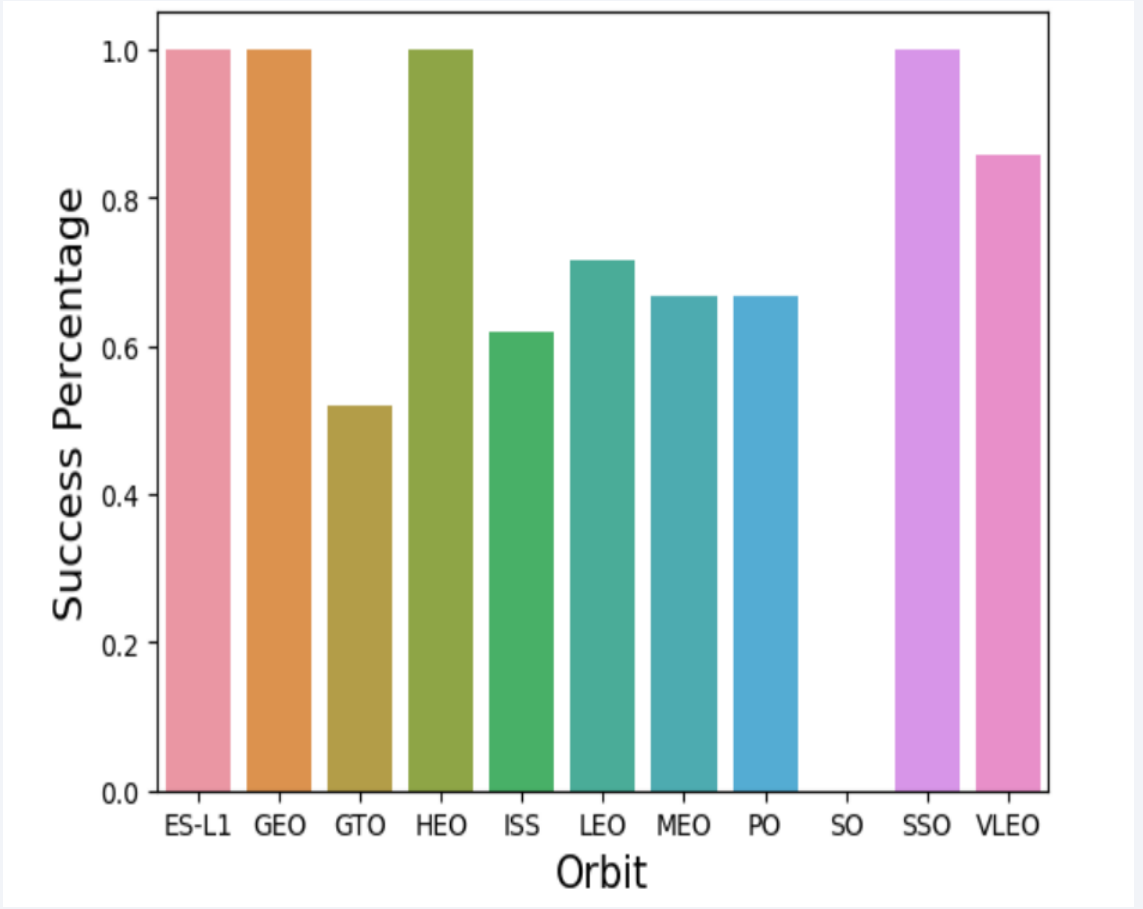
Payload vs. Orbit Type

- ISS orbit has the widest range of payload and a good rate of success;
- There are few launches to the orbits SO and GEO.
- VLEO only has high payload.



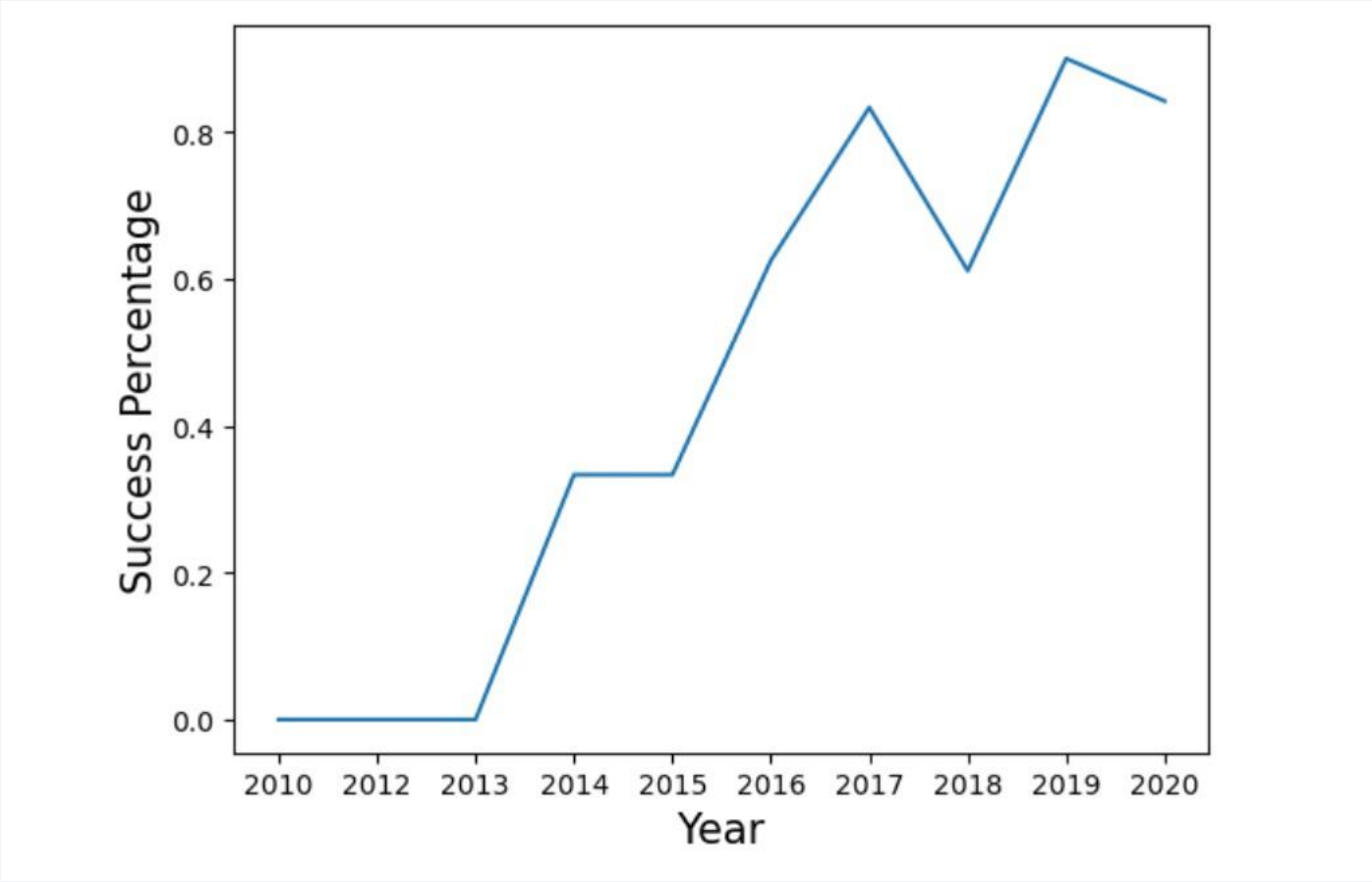
Success Rate vs. Orbit Type

- The biggest success rates happens to orbits:
  - ES L1;
  - GEO;
  - HEO; and
  - SSO.
- Followed by:
  - VLEO (above 80%); and
  - LFO (above 70%).



Launch Success Yearly Trend

- Success rate started increasing in 2013 and kept until 2020
- It seems that the first three years were a period of adjusts and improvement of technology.
- There was large dip in 2018 but it recovered in 2019



Average Payload Mass by F9 v1.1

- Filtering data by the booster version above and calculating the average payload mass we obtained the value of 2,928 kg.

1
2928

First Successful Ground Landing Date

- By filtering data by successful landing outcome on ground pad and getting the minimum value for date it's possible to identify the first occurrence, that happened on 12/22/2015.

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

- Selecting distinct booster versions according to the filters above, these 4 are the result.

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes

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

- Grouping mission outcomes and counting records for each group led us to the summary above.

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass ->
- We get this result by filtering data using a subquery to get max payload value.

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are

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

- We obtained this result by filtering the dataset on year - 2015 and outcome of failure.

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

- 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 ->

- We filtered the data on requested date and then grouped the filtered dataset by landing outcome and counted the no of entries in each group.

landing_outcome	COUNT
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


LAUNCH SITE
PROXIMITY ANALYSIS

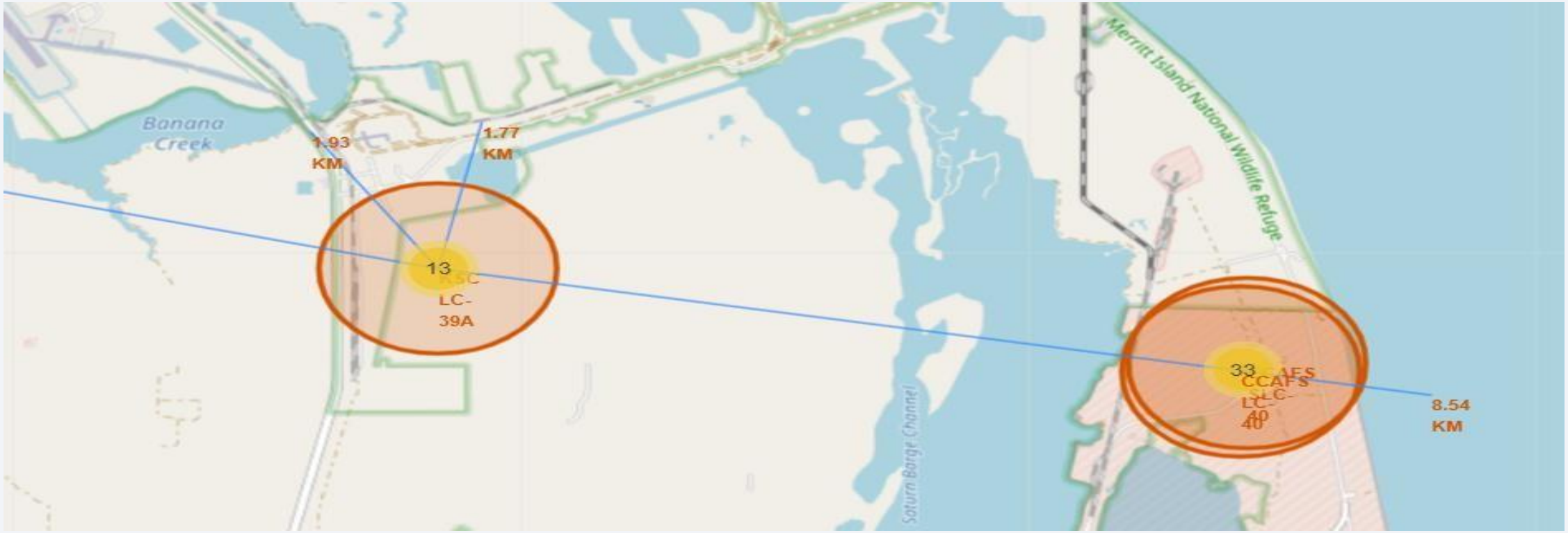
Launch Sites

- Launch sites are near sea, probably for safety reason, but not too far from roads and railroads.

Launch Outcomes by Site

- Above is Example of KSC LC 39A launch site launch outcomes.
- Green markers indicate successful and red ones indicate failure.

Logistics and Safety



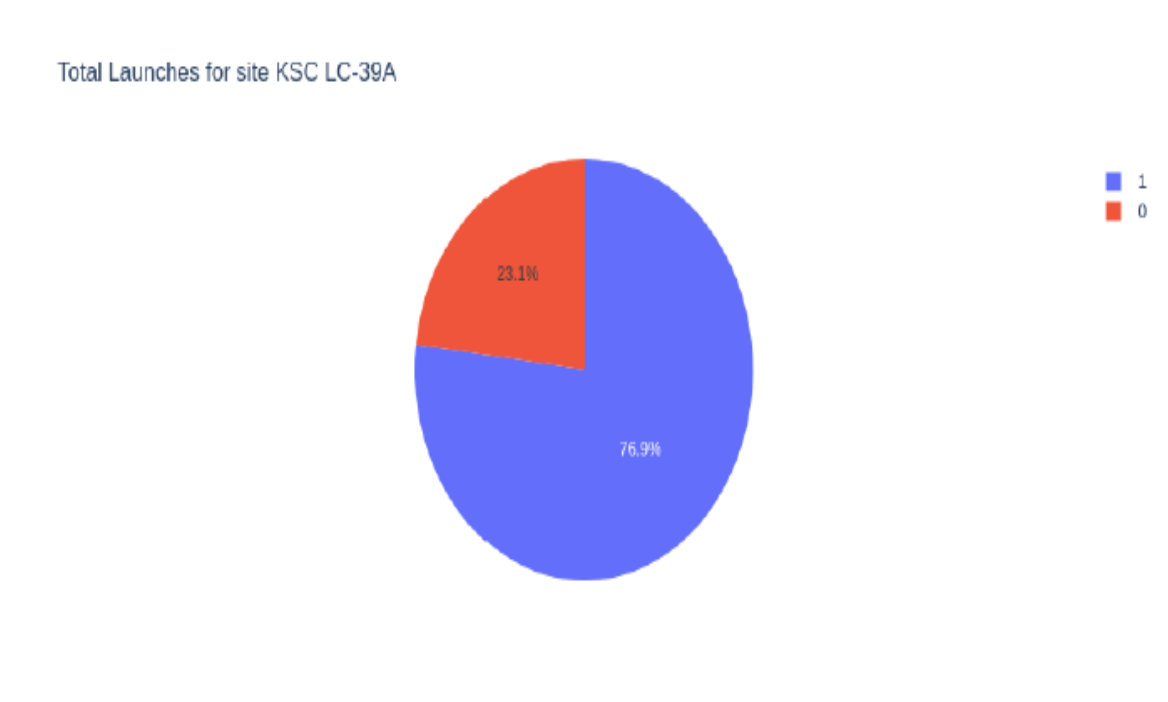
- Launch site KSC LC 39A has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

DASHBOARD

Successful Launches by Site

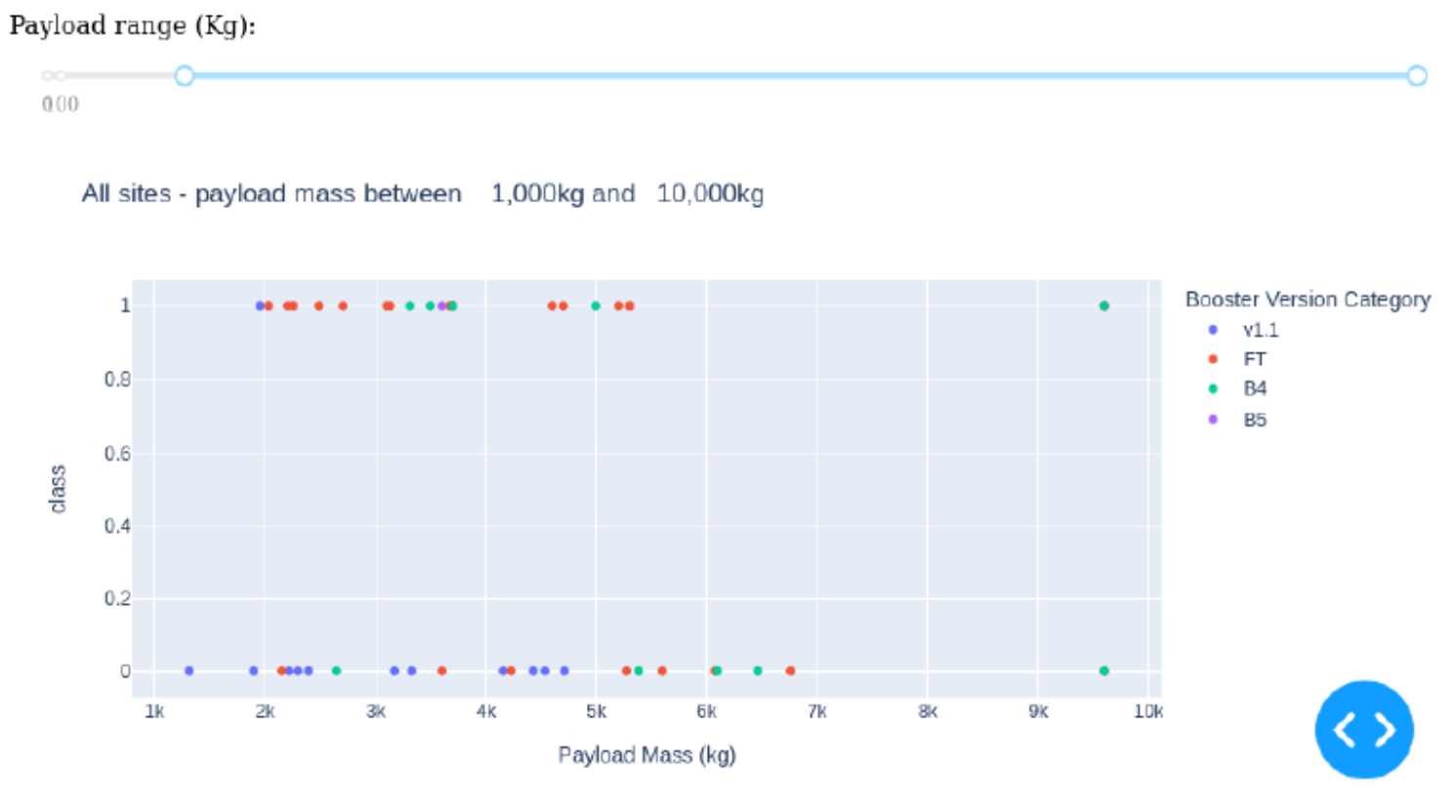
The place from where launches are done seems to be a very important factor of success of missions.

Launch Success Ratio for KSC LC-39A



- 76.9% of launches are successful in this site.

Payload vs. Launch Outcome



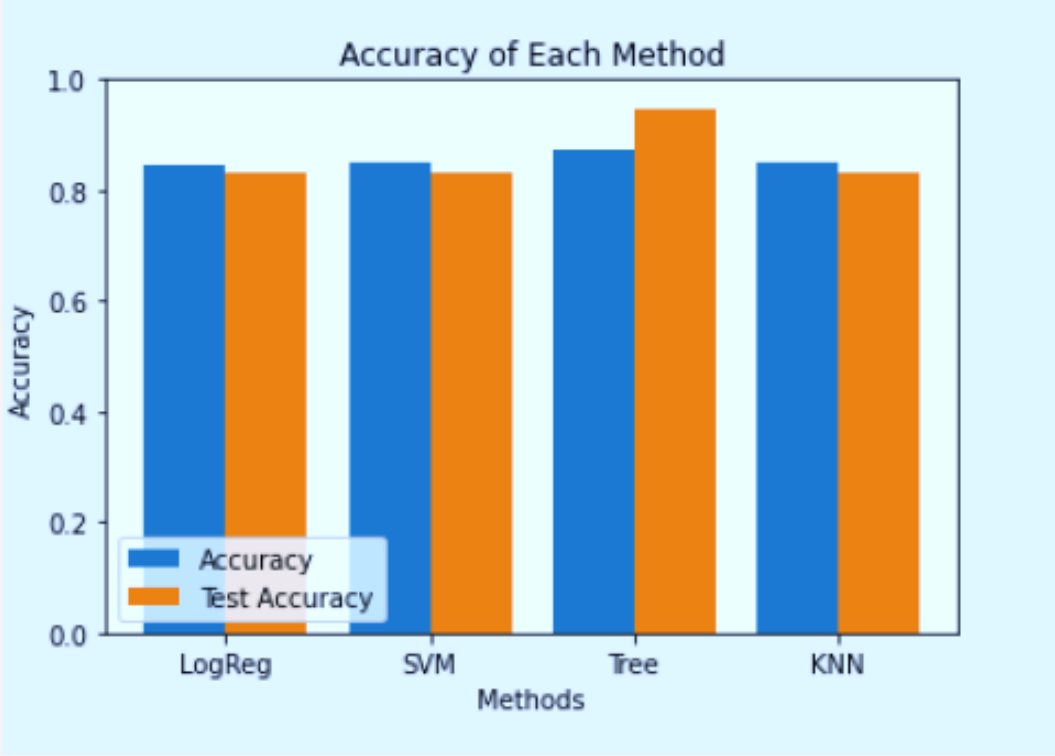
- Payloads under 6,000kg and FT boosters are the most successful combination.
- There's not enough data to estimate risk of launches over 7,000kg

Section 5

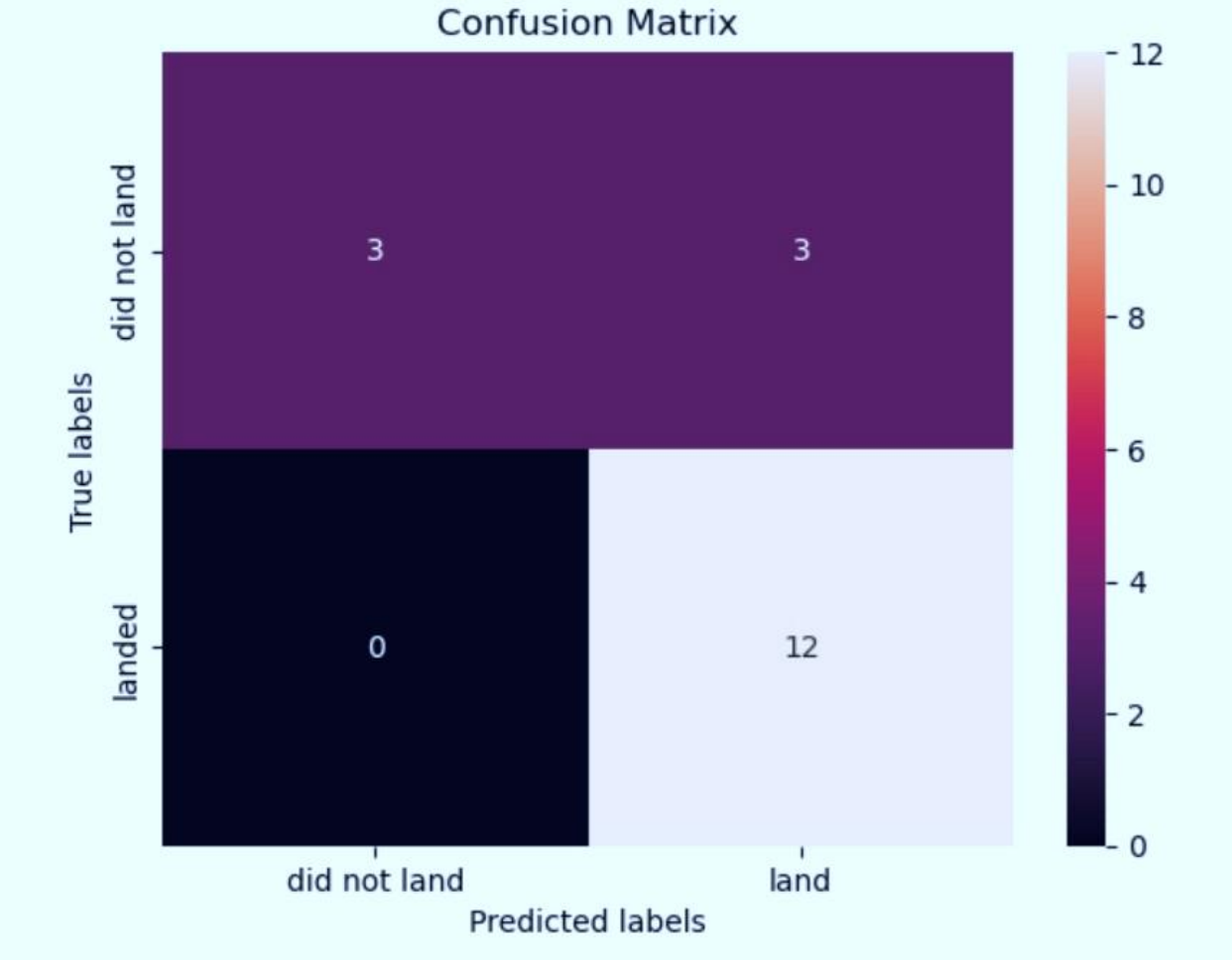
Predictive Analysis (Classification)

Classification Accuracy

- Four classification models were tested, and their accuracies are plotted beside
- The model with the highest classification accuracy is Decision Tree Classifier, which has accuracies over than 87%.



Confusion Matrix



- Confusion matrix of Decision Tree Classifier proves its accuracy by showing the big numbers of true positive and true negative compared to the false ones.

Conclusions

- Different data sources were analyzed, refining conclusions along the process;
- The best launch site is KSC LC 39A;
- Launches above 7,000kg are less risky;
- Although most of mission outcomes are successful, successful landing outcomes seem to improve over time, according the evolution of processes and rockets;
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Appendix

- Folium didn’t show maps on Github, so I took screenshots.
- For SQL queries output ss of jupyter notebook is attached

THANK YOU!