IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Emilia Dobkowska
08.02.2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies:
  - Data collection: SpaceX REST API and WebScraping from Wikipedia
  - Data wrangling (handling missing values, data cleaning and unification)
  - Exploratory Data Analysis (EDA) with visualisation (python matplotlib, seaborn) and SQL
  - Interactive visual analytics with Folium Map and Plotly Dash
  - Machine Learning predictive analysis using classification models (Logistic Regression, SVM, Decision Tree, KNN)
- Summary of all results:
  - It has been observed that success of the rocket first stage landing depends on many parameters such as: Flight Number, Launch Site, Booster version, Payload Mass, Orbit type, Year of launch
  - The highest prediction accuracy of successful landing has been obtained for Decision Tree classification model with accuracy 0.89 for test data

Introduction

- Project background and context:
  - The aim of the project to predict if the SpaceX Falcon 9 first stage will land successfully
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage
  - Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch
- We want to find answers for the following problems:
  - Which parameteres (of a rocket, launch site etc.) and how impact the probability of the first stage landing success?
  - For example does it depend on the launch site location, payload mass of the rocket, orbit type, flight number, booster version, …?
  - Which prediction model is the most accurate to predict future launch succes or failure?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Use of SpaceX REST API and webscraping from Wikipedia site
- Perform data wrangling
  - Preparation of data to facilitate later analysis (data cleaning and unification, selection of proper data from different sources, …)
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Accuracy comparison of four models: Logistic Regression, SVM, Decision Tree, KNN

Data Collection

- Two sources of data:
  - SpaceX REST API: https://api.spacexdata.com/v4/ to collect details about the rockets, launch sites and past launches
  - Web scrapping of Wikipedia site: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches to collect Falcon 9 historical launch records

Data Collection – SpaceX API

1 **REST API GET requests**

https://api.spacexdata.com/v4/

https://api.spacexdata.com/v4/rockets/	https://api.spacexdata.com/v4/launchpads/	https://api.spacexdata.com/v4/payloads/	https://api.spacexdata.com/v4/cores/	https://api.spacexdata.com/v4/launches/past
Booster names and versions	Launch sites: - Name - Latitiude - Longitute	- Payload mass - Orbit type	- Landing outcome and type - Number of flights with specific core - Gridfins used or not - Core reused or not - Number of times the core has been reused - Block of the core - Serial numer of the core - Legs used or not - Landing pad	Info about past launches

2 **Convert data into a Pandas data frame**

GitHub URL: https://github.com/uemk/IBM-Data-Science-Specialization/blob/master/Data%20Collection%20with%20API.ipynb

Data Collection - Scraping

With Python requests and Beautiful Soup libraries:

https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

1. Extract a Falcon 9 launch records HTML table from Wikipedia

   'Flight No.', 'Date and time', 'Launch site', 'Payload', 'Payload mass', 'Orbit', 'Customer', 'Launch outcome'

2. Parse the table and convert it into a Pandas data frame

GitHub URL: https://github.com/uemk/IBM-Data-Science-Specialization/blob/master/Data%20Collection%20with%20Web%20Scraping.ipynb

Data Wrangling

Identifiy and deal with missing values

- PayloadMass – replace with mean value
- LandingPad – leave None values to indicate pad was not used

First analysis of data to find some patterns and determine what would be the label for training supervised models

- Calculate the number of launches on each launch site
- Calculate the number and occurrence of each orbit
- Calculate the number and occurence of mission outcome per orbit type

Create a landing outcome label and add it to DataFrame

- New „Class” column – 0: failure, 1: success

Export new DataFrame to csv file

GitHub URL: https://github.com/uemk/IBM-Data-Science-Specialization/blob/master/Data%20Wrangling.ipynb

EDA with Data Visualization

- Plotted charts (each one vs outcome of the launch) :
  - Fligth Number vs Payload Mass
  - Flight Number vs Launch Site
  - Payload vs Launch Site
  - Success rate of each orbit
  - Flight Number vs Orbit type
  - Payload vs Orbit type
  - Launch success yearly trend

To better understand how different parameters such as: flight numer, payload, launch site, orbit type, year of launch impact the success of the launch

- GitHub URL: https://github.com/uemk/IBM-Data-Science-Specialization/blob/master/EDA%20with%20Visualisation.ipynb

EDA with SQL

- Summary of performed SQL queries:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date when the first successful landing outcome in ground pad was achieved
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - List the total number of successful and failure mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass
  - List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
  - Rank 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

- GitHub URL: https://github.com/uemk/IBM-Data-Science-Specialization/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Map objects created and added to a folium map:
  - Circles and markers for the all launch sites to display circle in an exact location of the site and its name as a popup label => to better understand which types of locations are chosen for launch sites (e.g. proximity of the ocean, cities, roads, etc.)
  - Marker cluster to mark the success/failed launches for each site on the map with different colours (red – failed, green – success)
  - Mouse Position to get coordinate for a mouse over a point on the map
  - Poly Lines and coresponding Markers to display the distance between selected launch site and coastline and other distances such as btw launch site and the closest city
- GitHub URL: https://github.com/uemk/IBM-Data-Science-Specialization/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Plots/graphs and interactions added to a dashboard:
  - Pie charts to visualise the percentage of success launches depending on the site:
    - total success launches by site
    - success vs failed launches per each site
  - Scatter plots to visualise dependancy of launch outcome from payload mass and booster version
  - Slider for payload range to dynamically visualise how payload impacts the success of the launch
- GitHub URL: https://github.com/uemk/IBM-Data-Science-Specialization/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

Standardize the data

↓

Split into training data and test data

↓

Find best Hyperparameter for SVM, Decision Tree, Logistic Regression and KNN

↓

Find the method performs best using test data

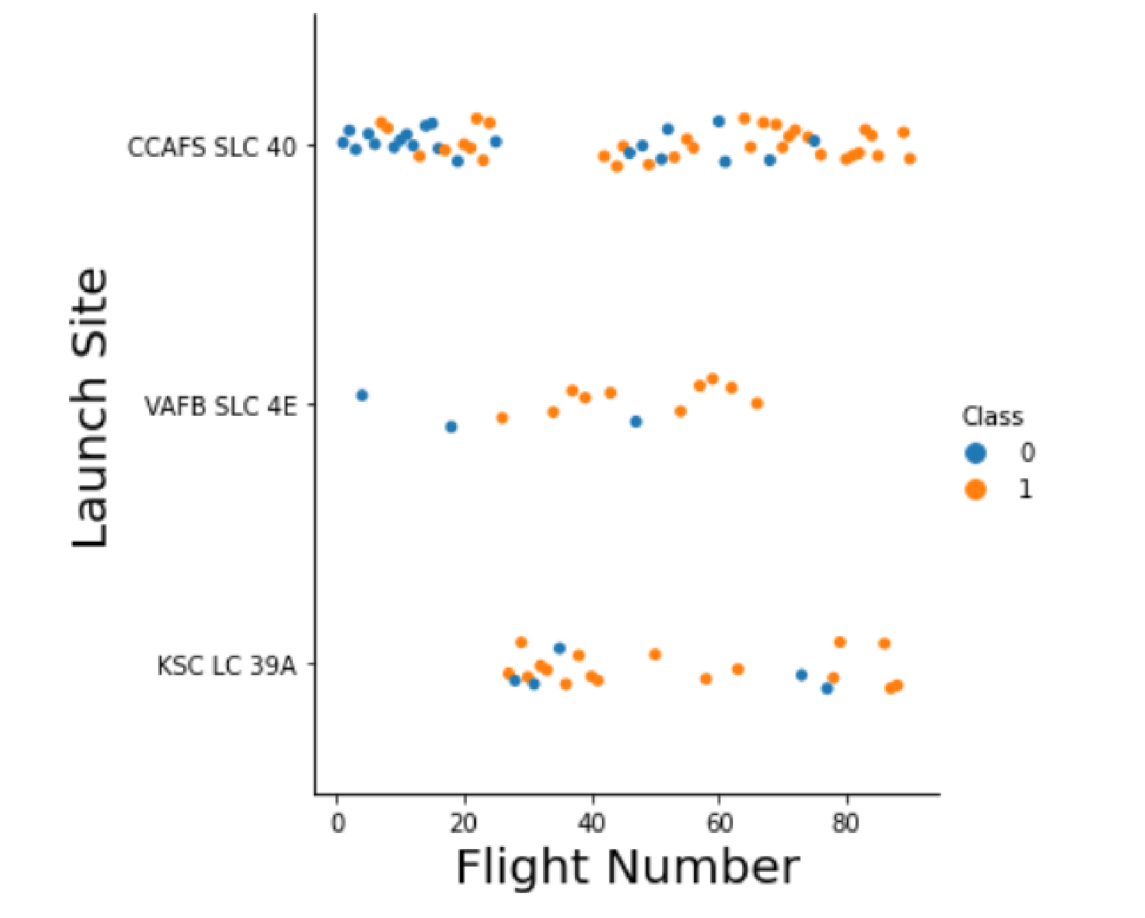
GitHub URL: https://github.com/uemk/IBM-Data-Science-Specialization/blob/master/Machine%20Learning%20Prediction.ipynb

Section 2

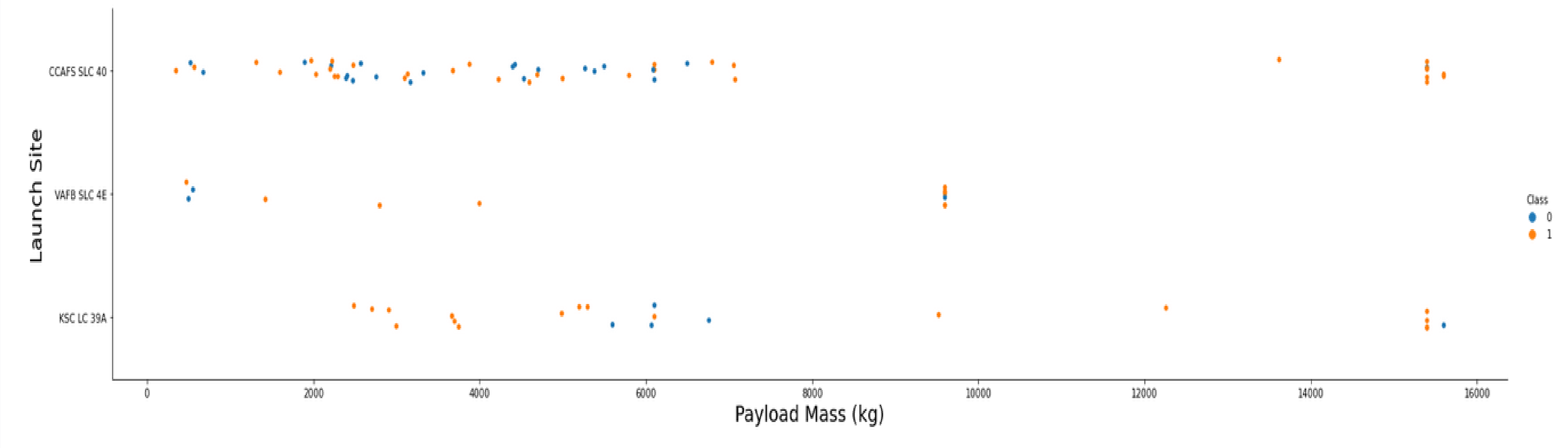
Insights drawn from EDA

Flight Number vs. Launch Site

- With the increase of the flight number we can observe that the probability of the launch success is generally higher
- This can be observed for all launch sites



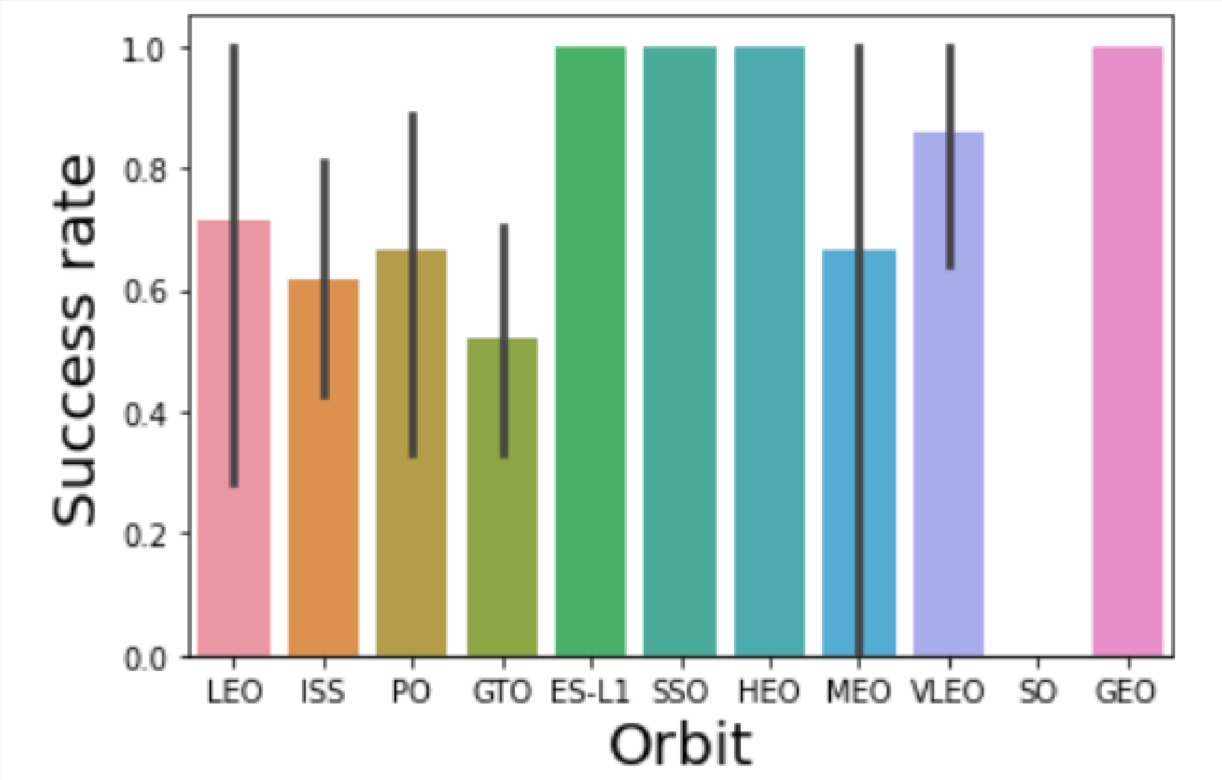
Payload vs. Launch Site



- For the VAFB-SLC launchsite there are no rockets launched for heavy payload mass(greater than 10000)
- Most of the launches are with lower payload mass

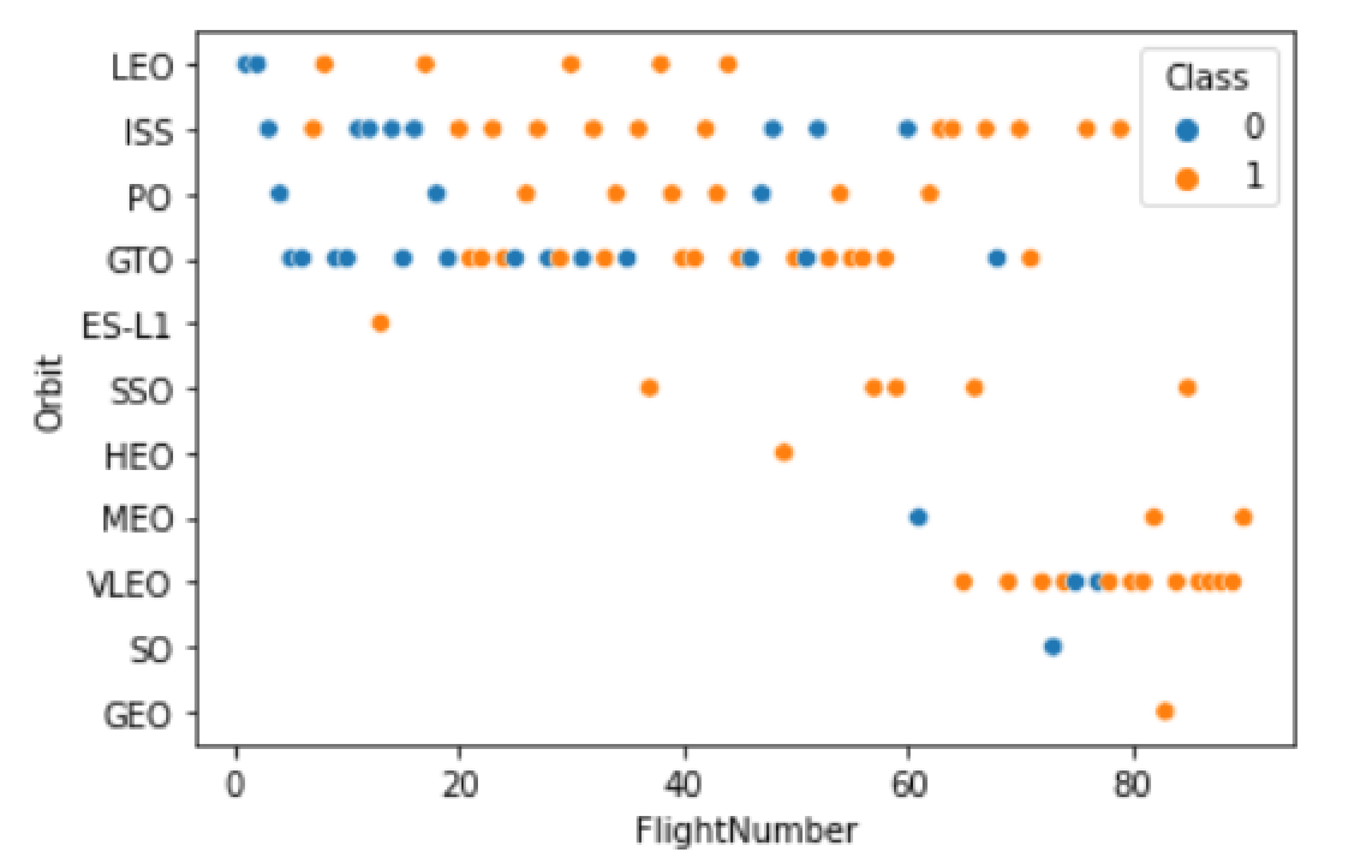
Success Rate vs. Orbit Type

- There are 4 orbit types with the highest launch success rate: ES-L1, SSO, HEO and GEO



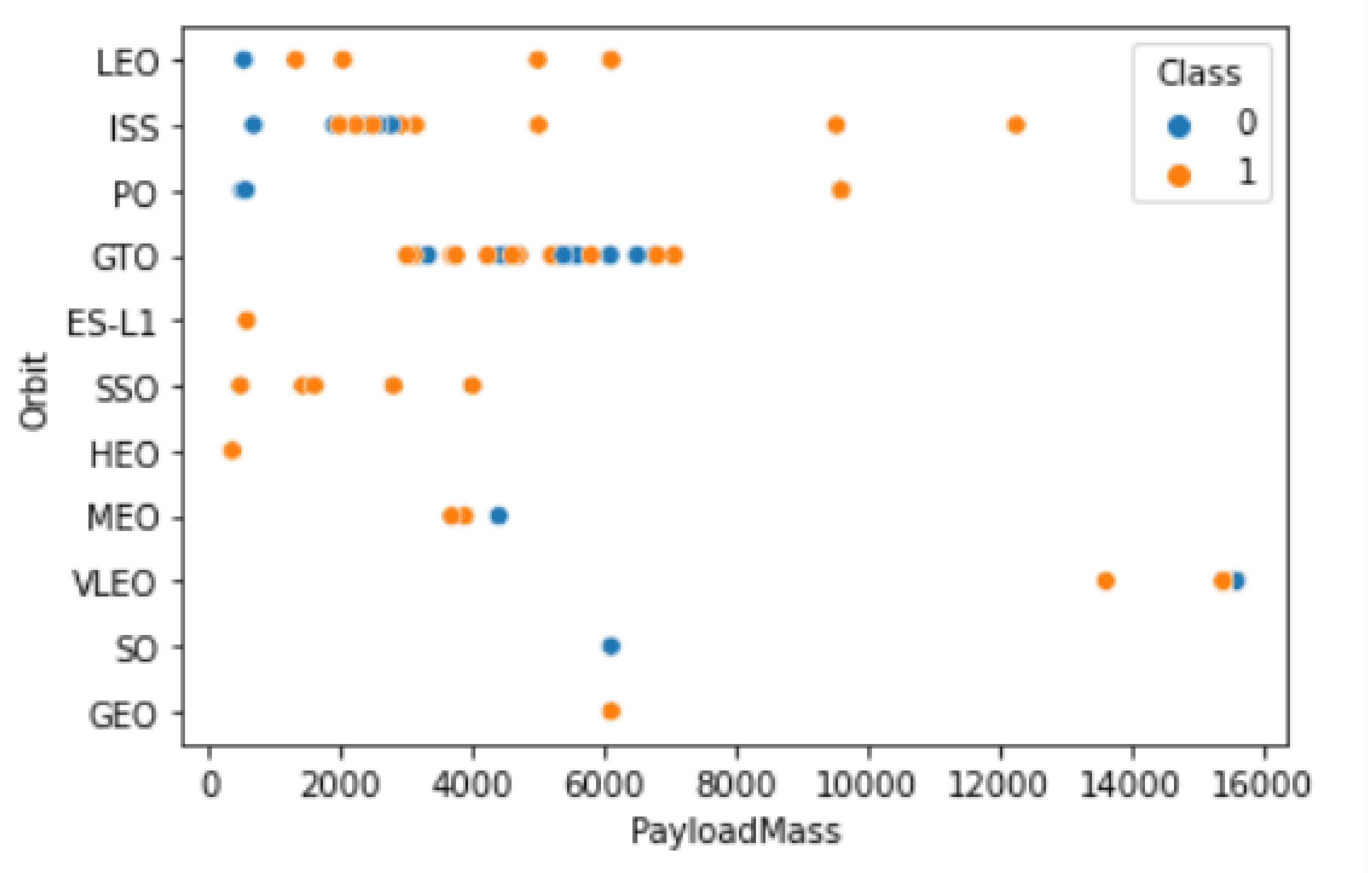
Flight Number vs. Orbit Type

- 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



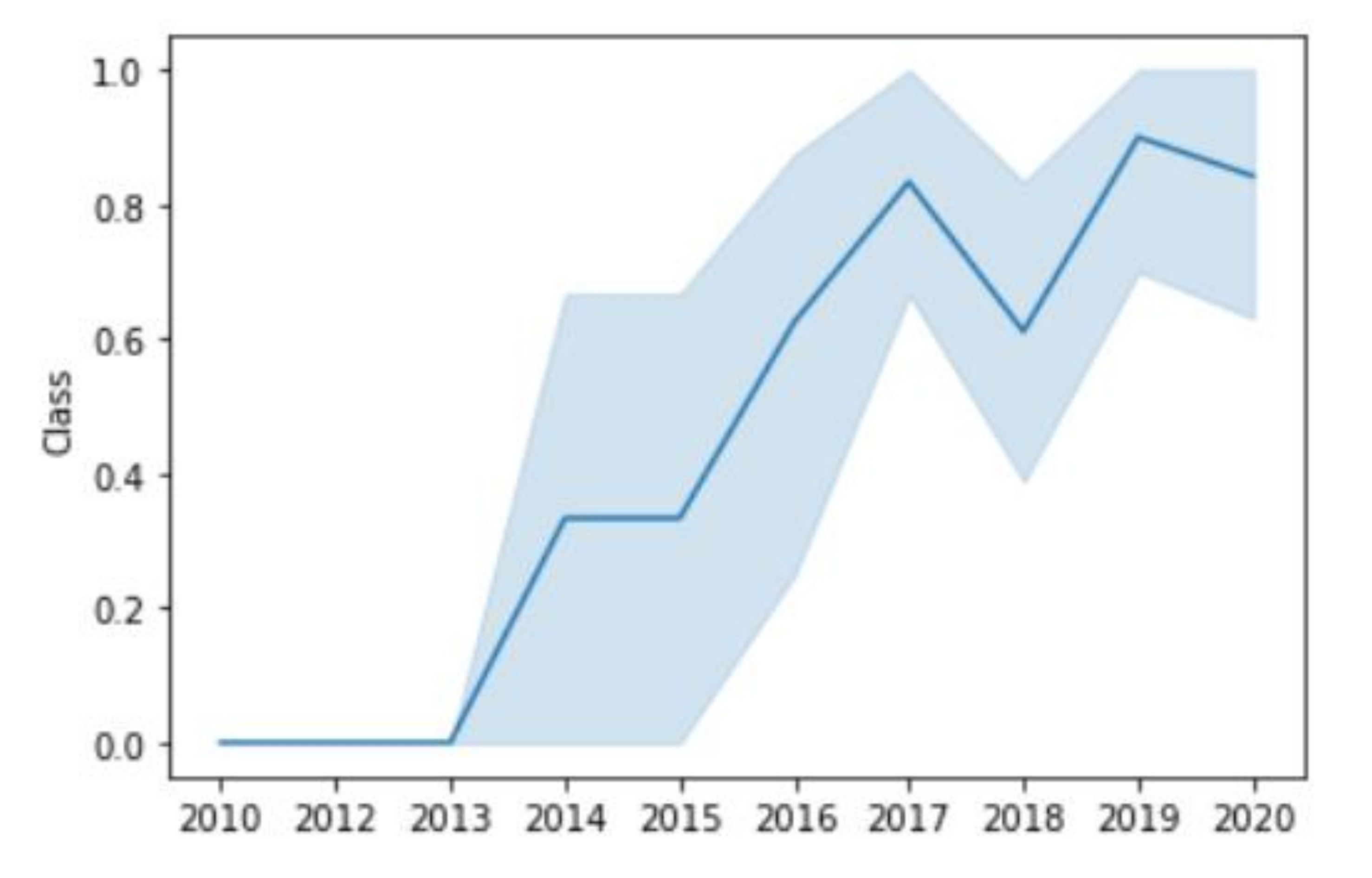
Payload vs. Orbit Type

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



Launch Success Yearly Trend

- The success rate since 2013 kept generally increasing till 2020



All Launch Site Names

- Unique launch sites retrieved with SQL query:

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA`:

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

- Total payload carried by boosters from NASA:

total_payload
45596 kg

Average Payload Mass by F9 v1.1

Average payload mass carried by booster version F9 v1.1:

avg_payload
2534 kg

First Successful Ground Landing Date

- First successful landing outcome on ground pad:

first_successful
2010-06-04

Successful Drone Ship Landing with Payload between 4000 and 6000

- 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

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes:

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

Boosters Carried Maximum Payload

- Names of the booster which have carried the 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

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015:

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

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

- Rank 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:

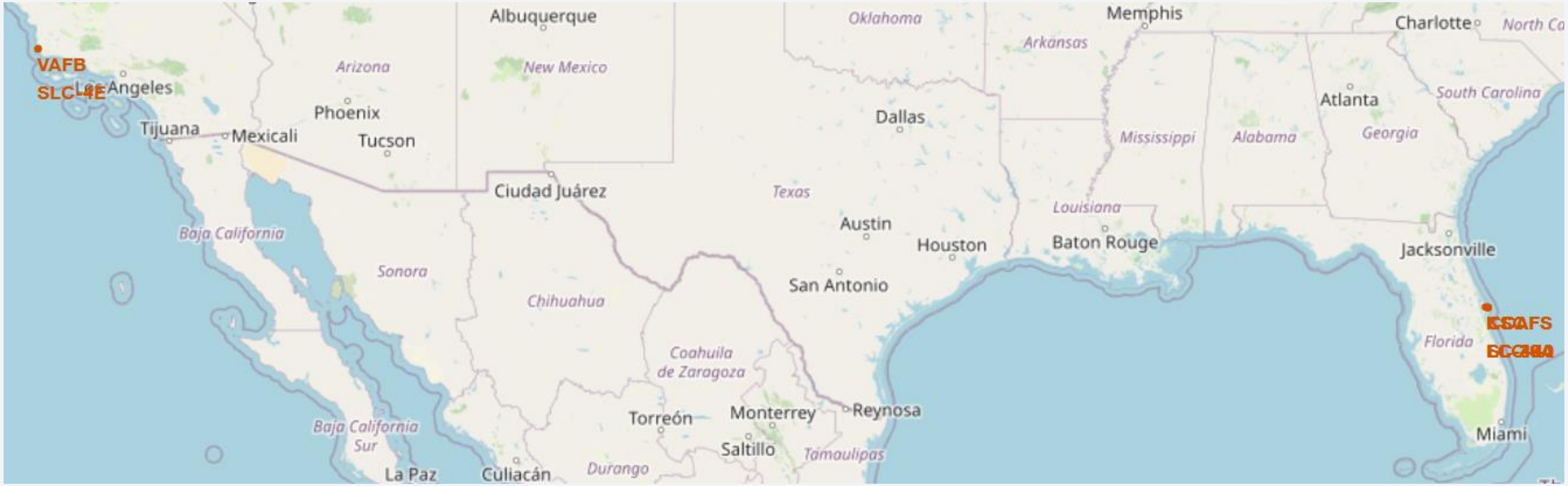
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

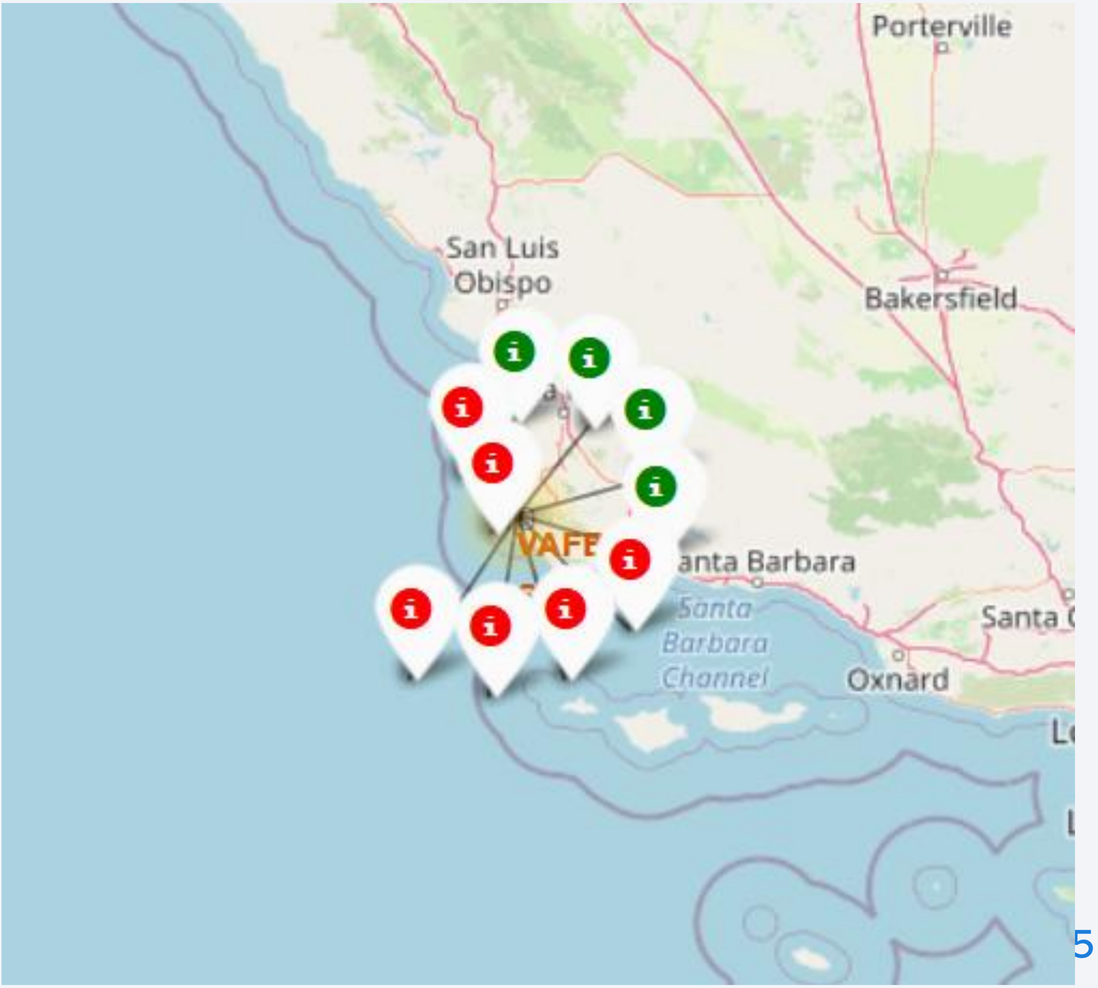
Folium Map – Launch Sites

- All Launch Sites are located close to the coast most probably to facilitate safe landing in the ocean



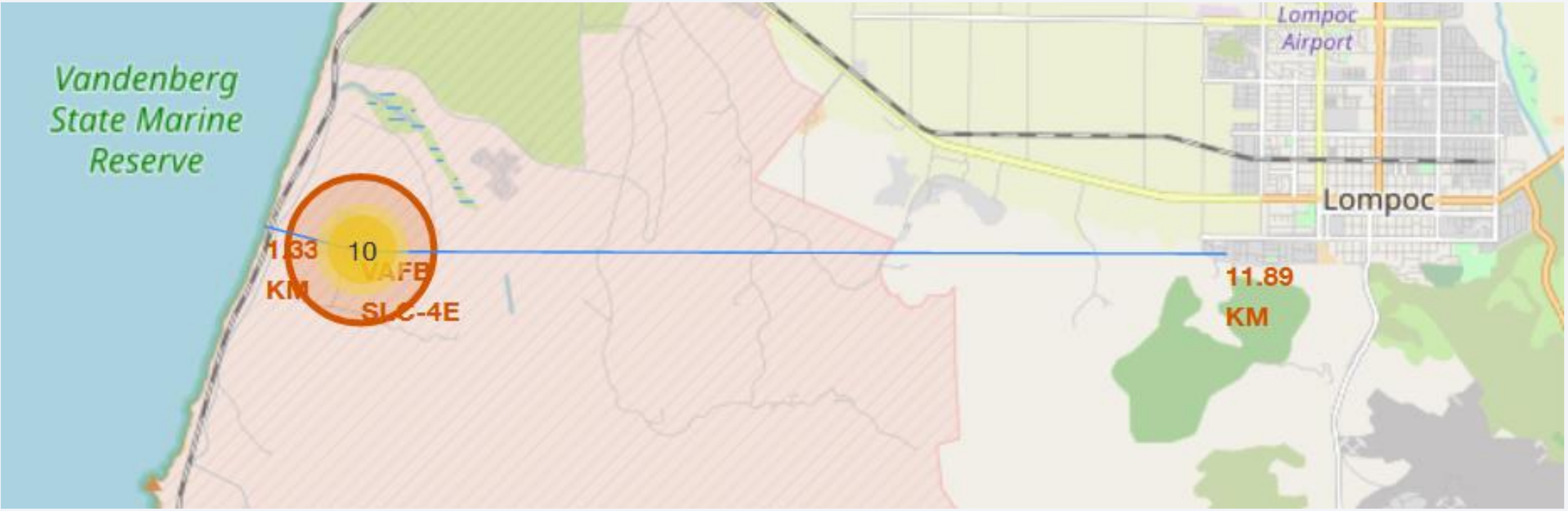
Folium Map – Launch Outcomes for VAFB SLC-4E

- From this visualisation we are able to identify which sites have relatively high success rates



Folium Map – Distance from VAFB SLC-4E to its proximities

- The selected site is located very close to the coast, it's also close to the railway
- However the distance to the closest city its more that 10km
- The distances to the public roads are also at least a few kms

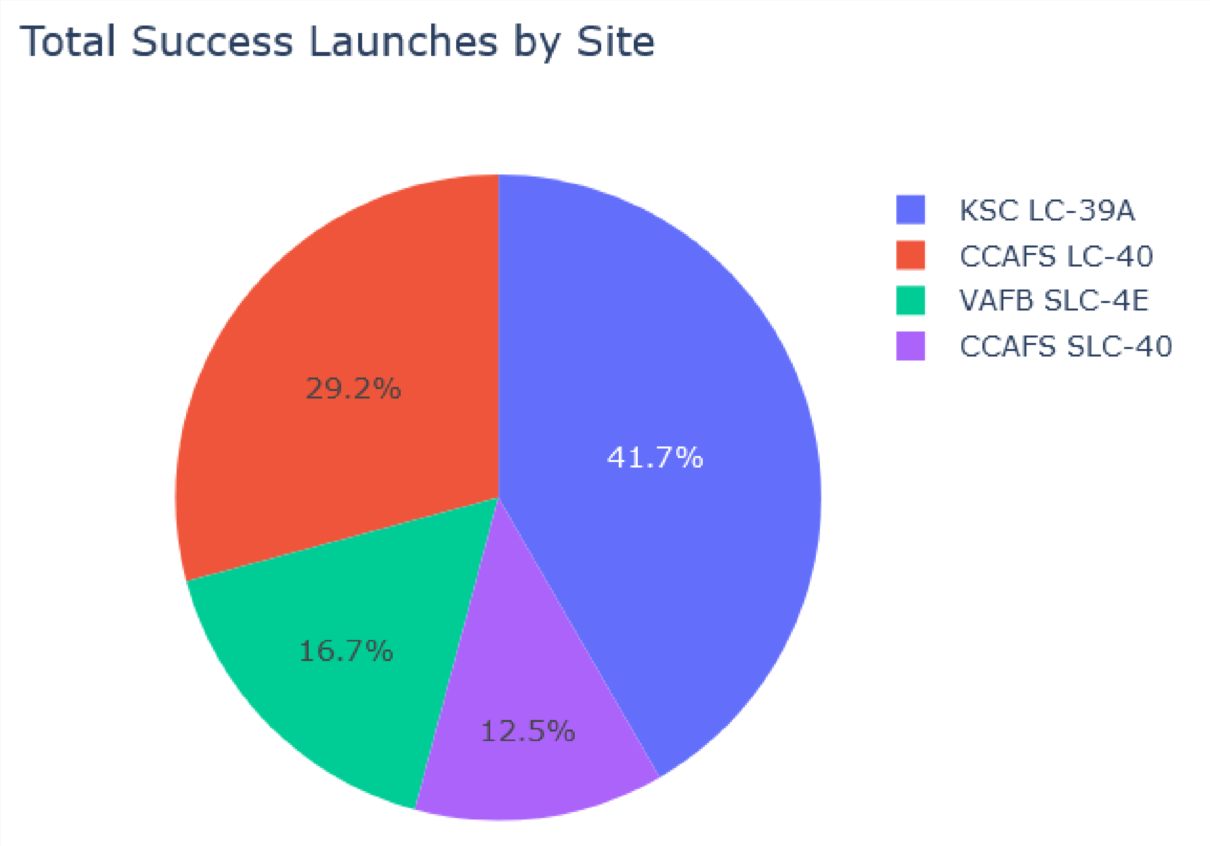


Section 4

Build a Dashboard with Plotly Dash

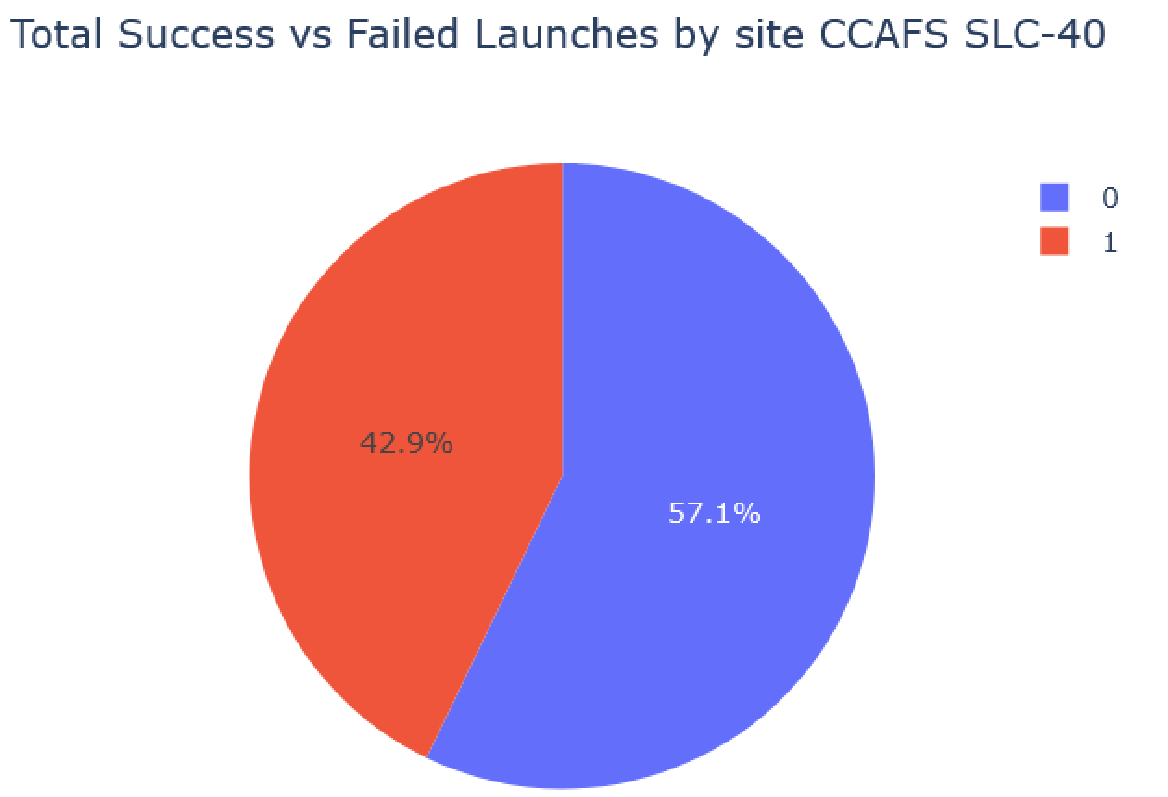
Plotly Dash – Total Success Launches by Site

- Among all sites the highest numer of successful launches was conducted from KSC LC-39A site, however, it may be due to the fact that the number of total launches is significantly higher from this site that others
- So it's also important to check success ratio per individual site



Plotly Dash – Success vs Failed Launches by site

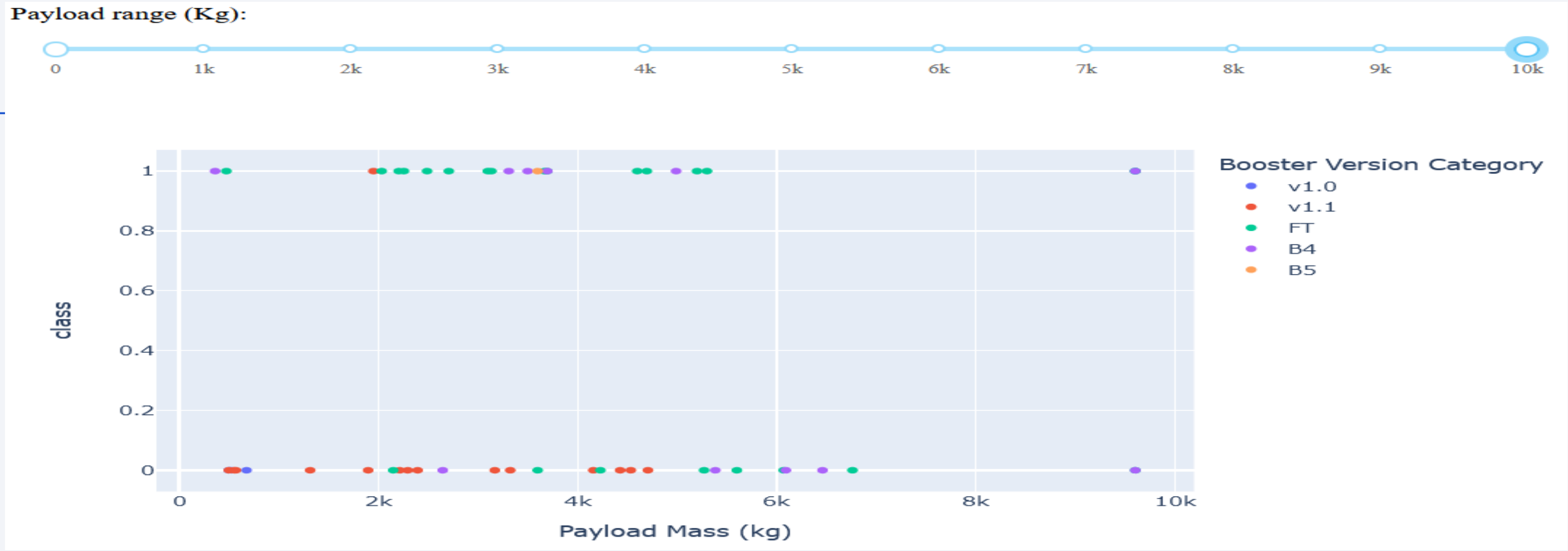
- The highest launch success ratio has CCAFS SLC-40 site: 42.9% of successful launches

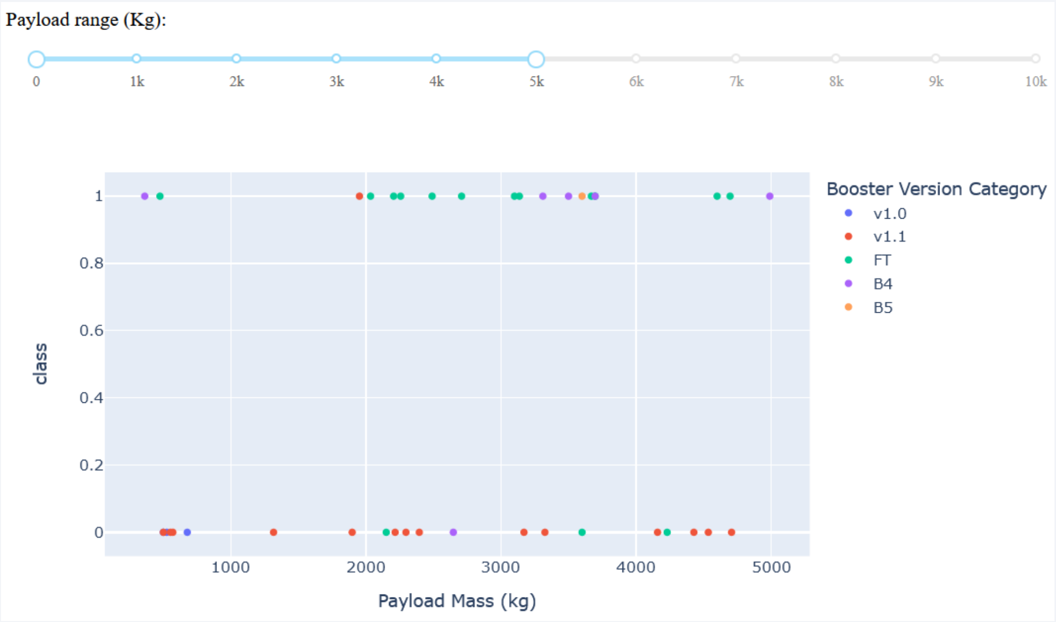


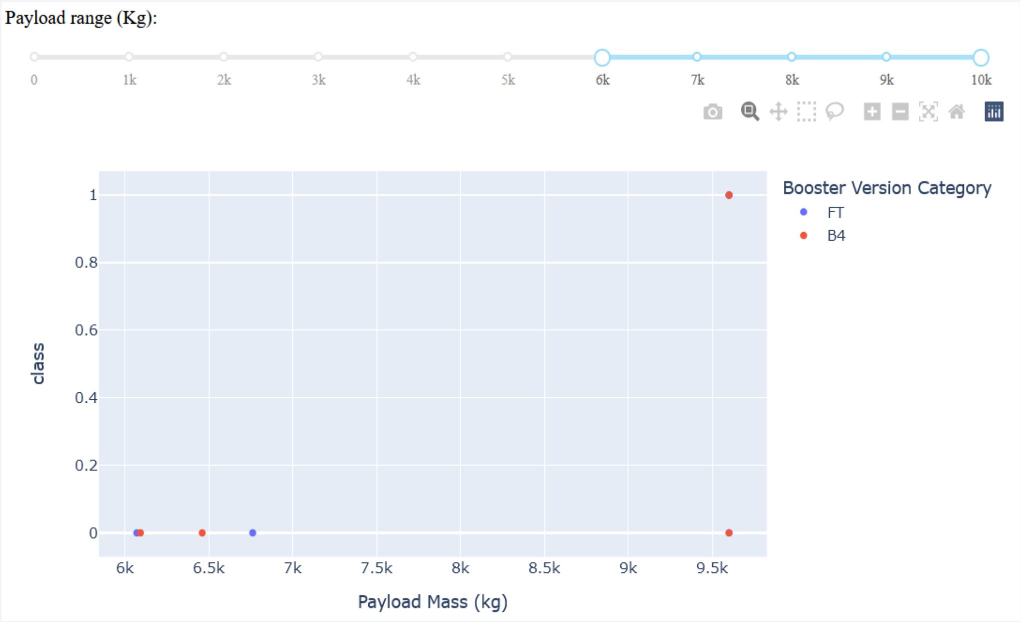
Plotly Dash – Payload vs Launch Outcome

The highest success rate is for:

- Booster versions: FT and B4 (the lowest for v1.1)
- Payload mass up to 5k kg





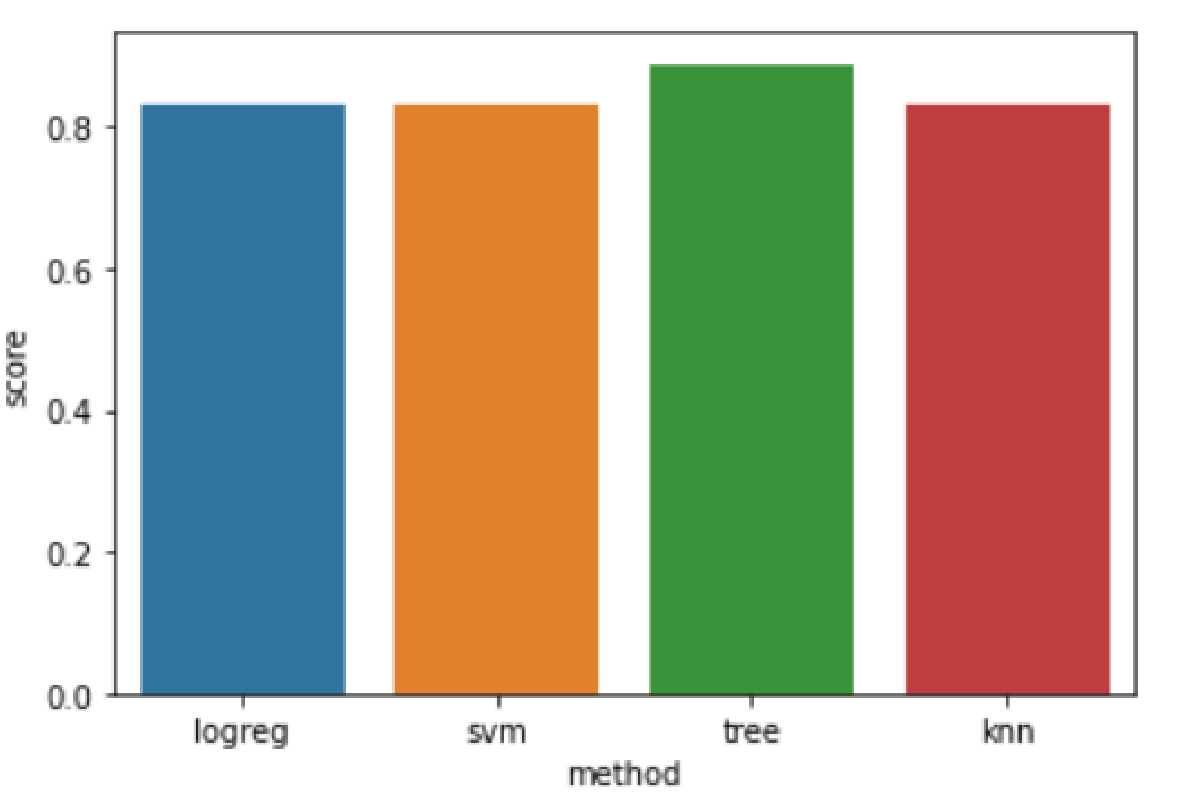


Section 5

Predictive Analysis (Classification)

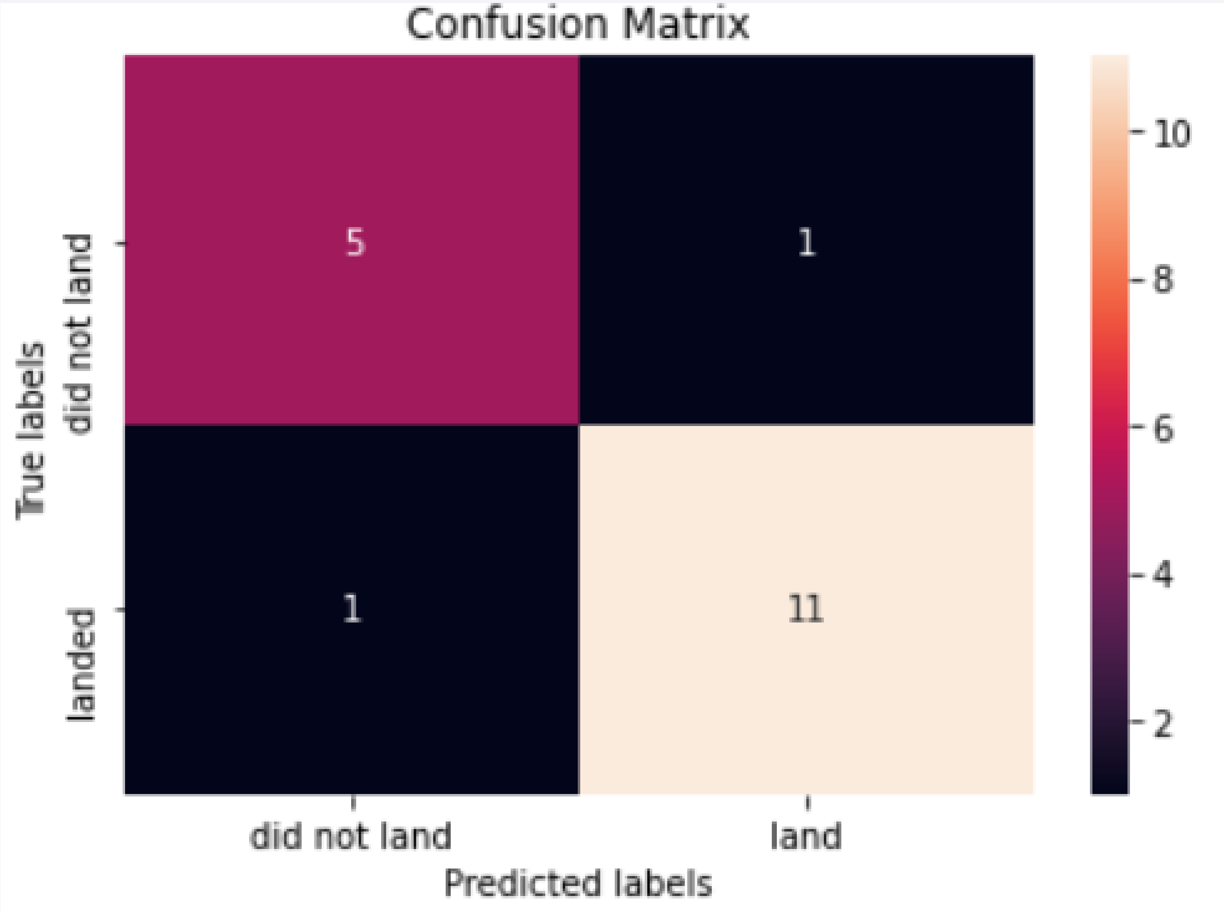
Classification Accuracy

- After comparison of accuracy of four classification models: SVM, Decision Tree, Logistic Regression and KNN the best results were obtained for **Decision Tree method**
- The obtained accuracy score for test data: 0.89



Confusion Matrix

- The confusion matrix obtained for Decision Tree model shows that the method quite well predicts both successful and failed launches (for test data)
- There is only 1 false positive and 1 false negative predicted label among total 18 samples



Conclusions

- It has been observed that success of the launch depends on many parameters such as:
  - Flight Number => the highest number the probability of success increases
  - Launch Site => different sites have different success vs failed launch ratio
  - Booster Category => highest success ratio for FT and B4 versions
  - Payload Mass => in general heavy payloads have lower success rates
  - Orbit type => 4 orbits with highest success ratio (ES-L1, SSO, HEO and GEO)
  - Year of launch => since 2013 increasing success rate
- However, this general observations are not allways true in every case: e.g. in general heavier payload results in lower success rate but when comparing it e.g. per orbit type this relation cannot be always confirmed
- The highest predition accuracy for test data (0.89) was obtained for Decision Tree model

Thank you!