IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Karolina Biernacka
20.03.2025

Outline

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

Executive Summary

- **The goal:** The project aim was to determine Space X launch price, through indicating features impacting SpaceX company stage landing success and predict the price of each launch
- **Methodology:** Data used included SpaceX and Wikipedia Falcon 9 launch information. Data was collected using REST API and web strapping, then joined and cleaned. Exploratory analysis was conducted using SQL, Seaborn visualizations, Folium map and dashboards. Finally the classification model predicting launch success was developed.
- **Results:** Landing success is highly related to payload mass, orbit, as well as launch site. Launch sites are located near equator and coastline to ensure efficiency and safety. The decision tree model can predict the success of first stage landing with 0.83 success rate.

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone. The most successful is SpaceX. One reason is that SpaceX rocket launches are relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upwards 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. Sometimes the first stage does not land. Sometimes it will crash as shown in this clip. Other times, Space X will sacrifice the first stage due to the mission parameters like payload, orbit, and customer.
- This is a data science project for a new rocket company. Space Y that would like to compete with SpaceX. The project aim is to determine features impacting first stage landing success and predict the price of each launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX launch data (SpaceX REST API) and Wikipedia web scraping
- Data wrangling
  - Convert various landing outcomes into categorical variable (Success/Not success)
- Exploratory data analysis (EDA) using visualization and SQL
- Interactive visual analytics using Folium and Plotly Dash
- Predictive analysis using classification models
  - Standarization, tunning and finding best performing model out of SVM, Classification Trees and Logistic Regression

Data Collection

Data Sources:

- SpaceX REST API including data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Wiki pages web scrapping to gather HTML tables that contain valuable Falcon 9 launch records using BeautifulSoup.

Raw data was transformed (todataframe format, tables joined and filtred, nulls completed) into a clean dataset which provides meaningful data on the Falcon 9 Launches.

Data Collection – SpaceX API

Request from SpaceX API

Normalize JSON file

Integrading data from API (ID, launches, booster version) into one dataframe

Filter booster version to include Falcon 9 only

Replacing missing pay load mass data with the mean

GitHub URL:
https://github.com/biernackakarolina/Space_Project/blob/58b713c3b4835467d383321faf5f0dbb1599cba3/01_jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping

Request the Falcon9 Launch Wiki page from its URL

Creating BeutifulSoup object

Extracting variable names from the HTML table header

Parsing the launch HTML tables into DataFrame

GitHub URL:
https://github.com/biernackakarolina/Space_Project/blob/58b713c3b4835467d383321faf5f0dbb1599cba3/02_jupyter-labs-webscraping.ipynb

Data Wrangling

Data wrangling process:

Checking data types and missing values	Calculating the number of launches on Space X launch sites and orbites, as well as mission outcomes on different orbits	Creating a landing outcome label from Outcome column If "True" (Success) then 1 If "False"/"None (Fail) then 0	Detemining the **overall landing outcome rate (0.67)**

GitHub URL:
https://github.com/biernackakarolina/Space_Project/blob/58b713c3b4835467d383321faf5f0dbb1599cba3/03_labs-jupyter-spacex-Data%20wrangling-v2.ipynb

EDA with Data Visualization

Data was explored with the following charts:

- Scatter plots visualizing relationship between:
    - flight number, pay load mass and landing success
    - flight number, launch site and landing success
    - load pass, launch site and landing success
    - light number and orbit type and landing success
    - pay load mass, orbit type and landing success
- Bar chart visualizing relationship between success rate of each orbit type
- Line plot visualizing launch success yearly trend

Based on the plots, several features were chosen for model predicting landing success. Numerical data were converted to 'float' type and categorical one-hot encoded.

GitHub URL: https://github.com/biernackakarolina/Space_Project/blob/58b713c3b4835467d383321faf5f0dbb1599cba3/05_jupyter-labs-eda-dataviz-v2.ipynb

EDA with SQL

Exploratory SQL querries performed:

- Selecting unique launch sites in the space mission
- Selecting 5 records where launch sites begin with the string 'CCA'
- Selecting total payload mass carried by boosters launched by NASA (CRS)
- Selecting average payload mass carried by booster version F9 v1.1
- Identifying the date when the first succesful landing outcome in ground pad was acheived
- Listing the names of the boosters which have success in drone ship and have payload between 4000 and 6000
- Listing the total number of successful and failure mission outcomes
- Listing the names of the booster_versions which have carried the maximum payload mass
- Ranking the count of landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

GitHub URL: https://github.com/biernackakarolina/Space_Project/blob/0472e02bd91a8a8736d93400640008253f0850d0/04_jupyter-labs-eda-sql-coursera_sqllite%20(1).ipynb

Build an Interactive Map with Folium

Map object added to a map:

- Launch sites circles and marks with pop up labels
- Success(green)/failed(red) launches for each site circles and marks
- Calculated distances between a launch site to its proximities with polylines

Objects were added to visualize where SpaceX launch sites were located (Landing success is highly related to payload mass, orbit, and launch site and those features interact) and what is success rate of landing in each site.

GitHub URL: https://github.com/biernackakarolina/Space_Project/blob/58b713c3b4835467d383321faf5f0dbb1599cba3/06_lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
- Explain why you added those plots and interactions
- Add the GitHub URL of your completed Plotly Dash lab, as an external reference and peer-review purpose

Predictive Analysis (Classification)

Models checked included SVM, Classification Tree, KNN, and Logistic Regression. Best performing classification model was found through the following steps:

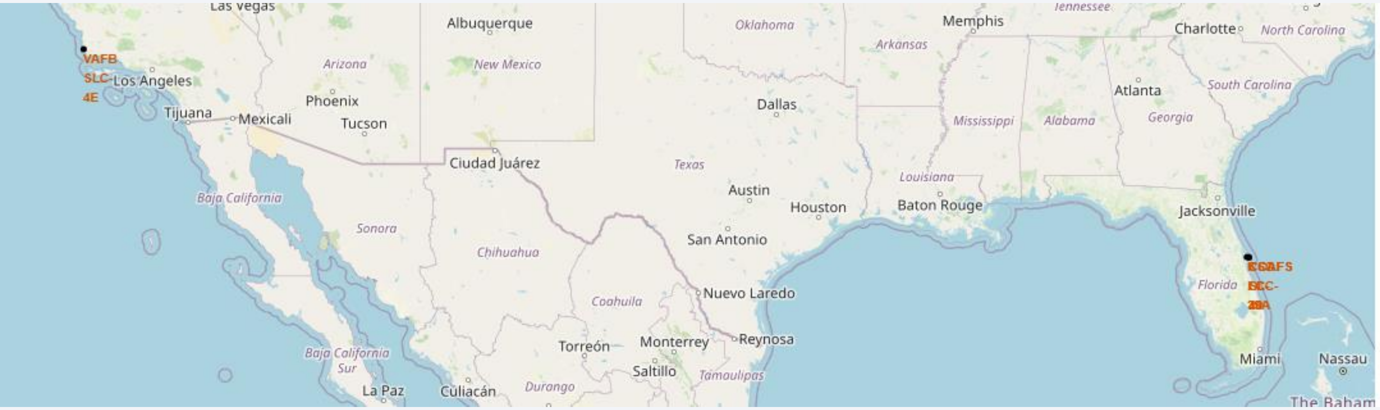
1. Splitting data into target and features:
   - Y – Class (Success-1/ Fail-0)
   - X – stardardized features using StandardScaler()
2. Splitting data into train and test sets using train_test_split
3. Finding best model parameters using GridSearchCV
4. Models evaluation through:
   - Train data accuracy score
   - Test data accuracy score
   - Confusion matrix
5. Findings:
   All models performed the same on the test set (0.83)

Although all models performed the same on the test set, the decision tree clacification model overperformed other models on the train set (0.87). It might be due to overfitting and it would be usefull to implement strateggies to overcome overfitting and explore this model more deeply

GitHub URL: https://github.com/biernackakarolina/Space_Project/blob/58b713c3b4835467d383321faf5f0dbb1599cba3/07_SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

- **Exploratory data analysis results** – the following features are relevant and impact first stage landing success: flight number, payload mass, orbit, launch site, number of flights, grid fins, whether stage was already reused and how many times, legs, landing pad, block, and serial
- **Interactive analytics demo in screenshots**

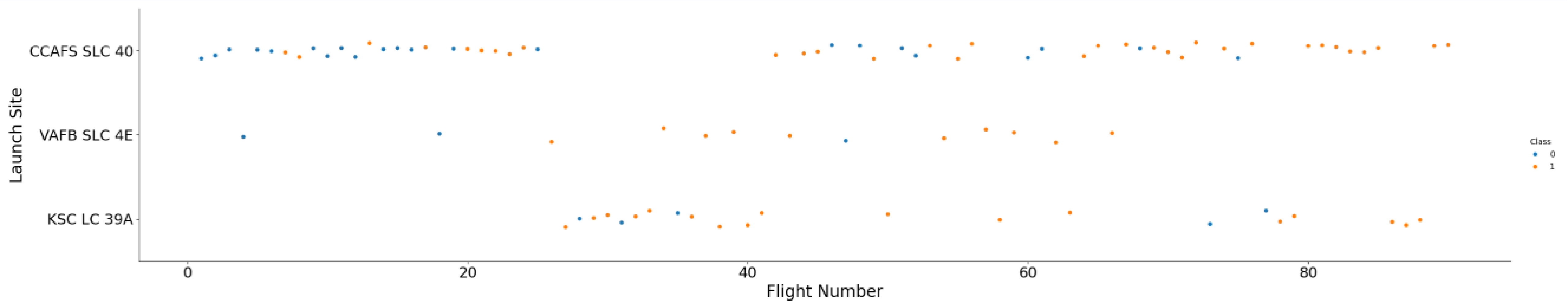


- **Predictive analysis results** – all evaluated classification models can predict the landing outcome with 0.83 accuracy score

Section 2

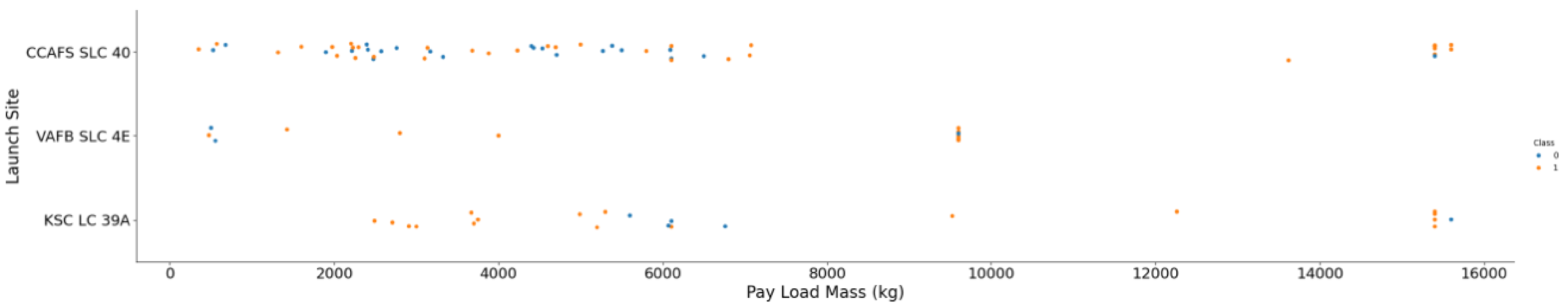
Insights drawn from EDA

Flight Number vs. Launch Site



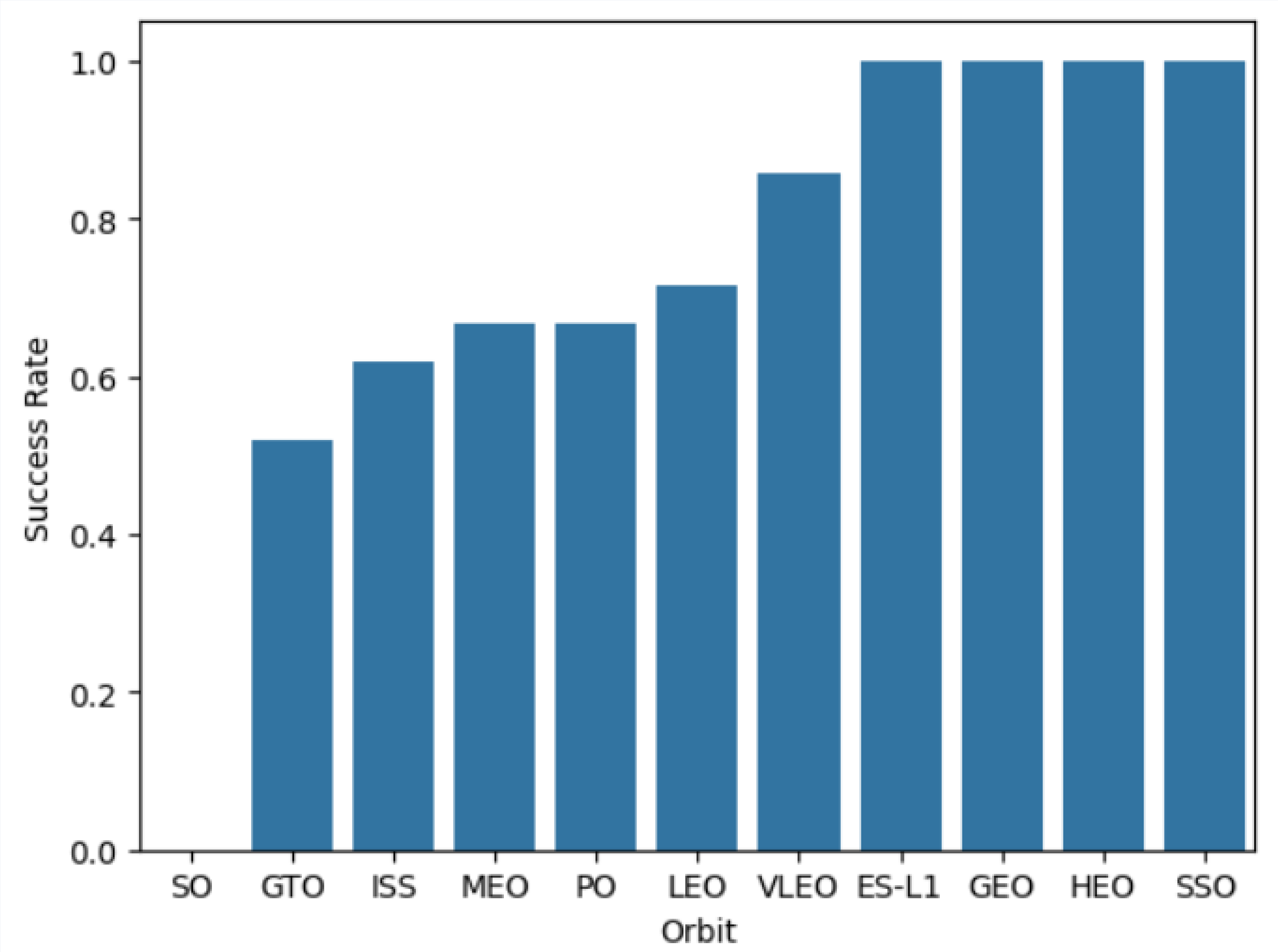
First launches were conducted from CCAFS SLC 40 launch site and back then landing failed more often (probably due to lack of experience and testing). The higher light number the higher success rate for all launch sites, with VAFB SLC 4E being the least common used but most reliable.

Payload vs. Launch Site



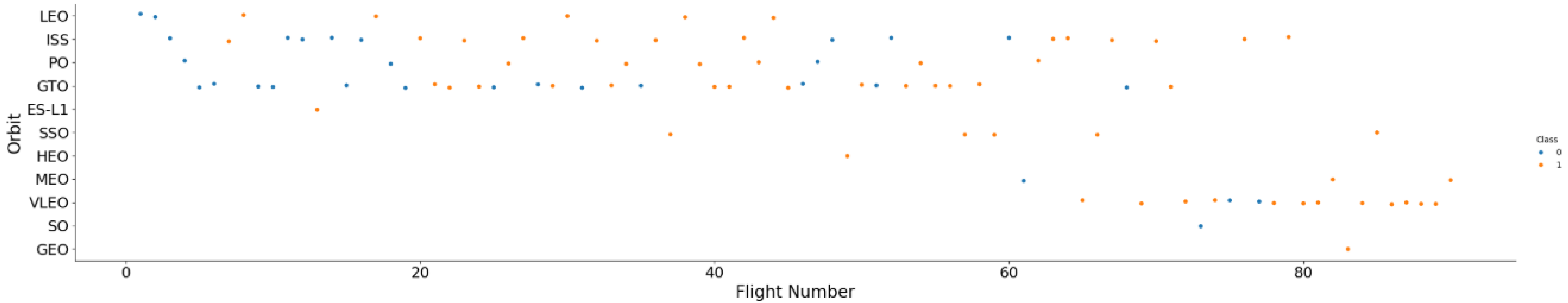
Lighter payloads are more common. For VAFB SLC 4E and KSC LC 39A, between 1000 and 5,500 payload all first stages returned succesfully.

Success Rate vs. Orbit Type



The success rate is the highest for VLEO, ES-L1, GEO, HEO and SSO (almost 100%), whereas the lowest for SO (non successful landing) and GTO (around 50% success).

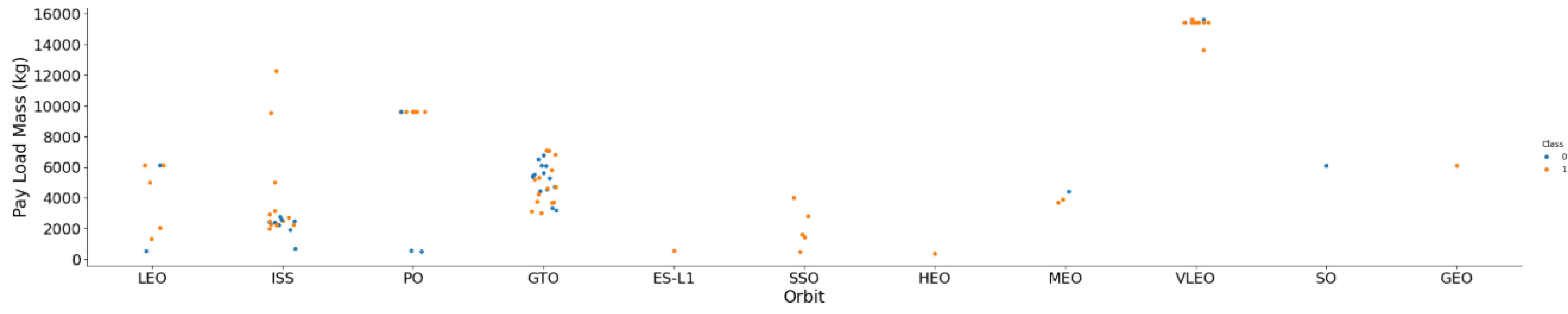
Flight Number vs. Orbit Type



First launches were conducted on LEO, ISS, PO and GTO orbits and back then landing failed more often (probably due to lack of experience and testing). Later launches were more successful and mostly conducted on ISS and VLEO.

Some of orbits with high success rate (ES-L1, GEO, HEO) had only one launch (not enough to indicate rate) whereas SSO and VLEO have high success rate with at least few launches completed.

Payload vs. Orbit Type



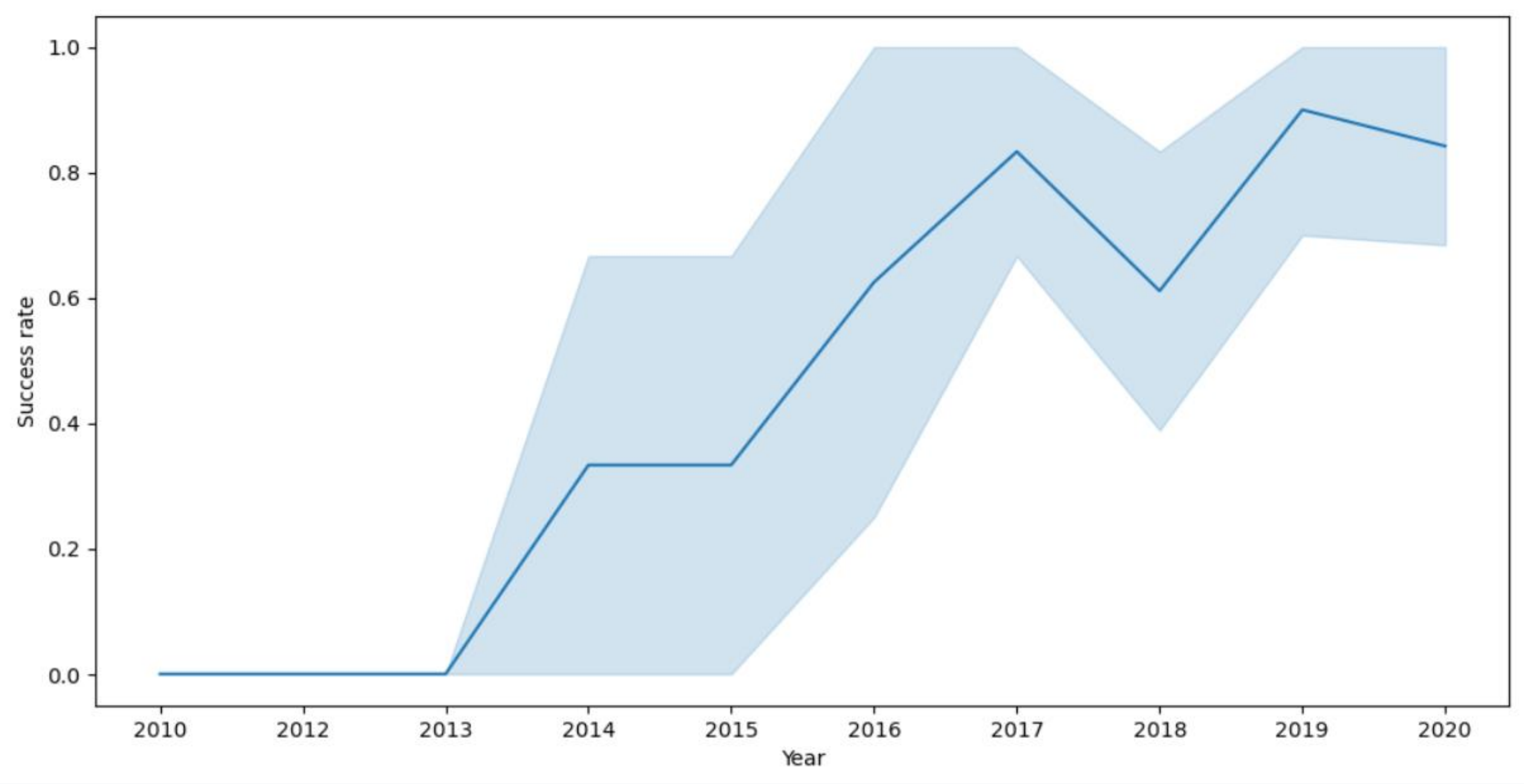
Rockets with different payload mass are sent to different orbits. For VLEO dominate high payload mass, whereas for LEO, ISS, GTO and SSO low-to-moderate payload mass. Almost all rockets sent to PO had 10,000 payload mass.

Success rate differs as well – it is high for VLEO where all rockets sent had high mass, it is high for SSO where all rockets sent had low mass, whereas it seems to be better for higher mass than for lower where various-mass rockets were sent (ISS, PO).

Launch Success Yearly Trend

Success rate rises with years.

This trend is probably a result of experience and testing, thus developing good launches strategy with time.



All Launch Site Names

The SpaceX launch sites are:

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

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

For first 5 records for CCAFS first stage landed successfully.

Total Payload Mass

Total payload carried by boosters from NASA is:

45,596 kg

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is

2,534.67 kg

First Successful Ground Landing Date

The first successful landing outcome on ground pad took place on:

22.12.2015

Successful Drone Ship Landing with Payload between 4000 and 6000

Boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are:

F9 FT B1020

F9 FT B1022

F9 FT B1026

F9 FT B1021.2

F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes*:

Success	Failure
100	1

*Those are statistic for the mission outcome, not first stage landing outcome being the main interest in this project

Boosters Carried Maximum Payload

Booster which have carried the maximum payload mass are:

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

Failure landing outcomes in drone ship 2015

month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

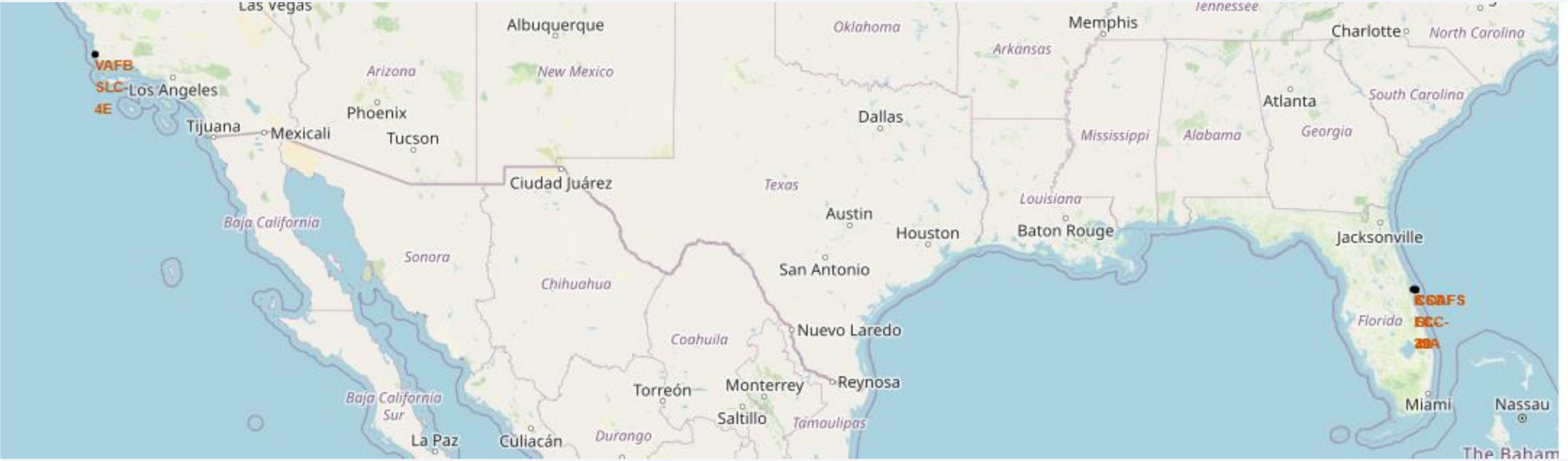
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

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

Section 3

Launch Sites Proximities Analysis

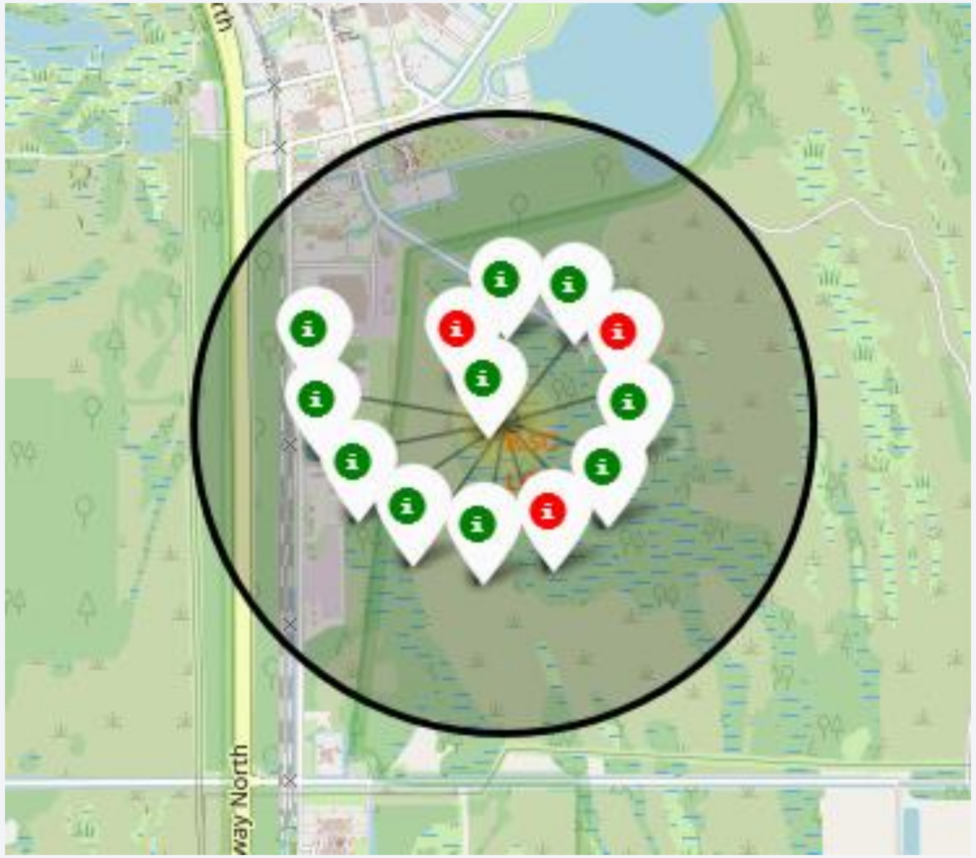
SpaceX Launch Sites Locations - Folium



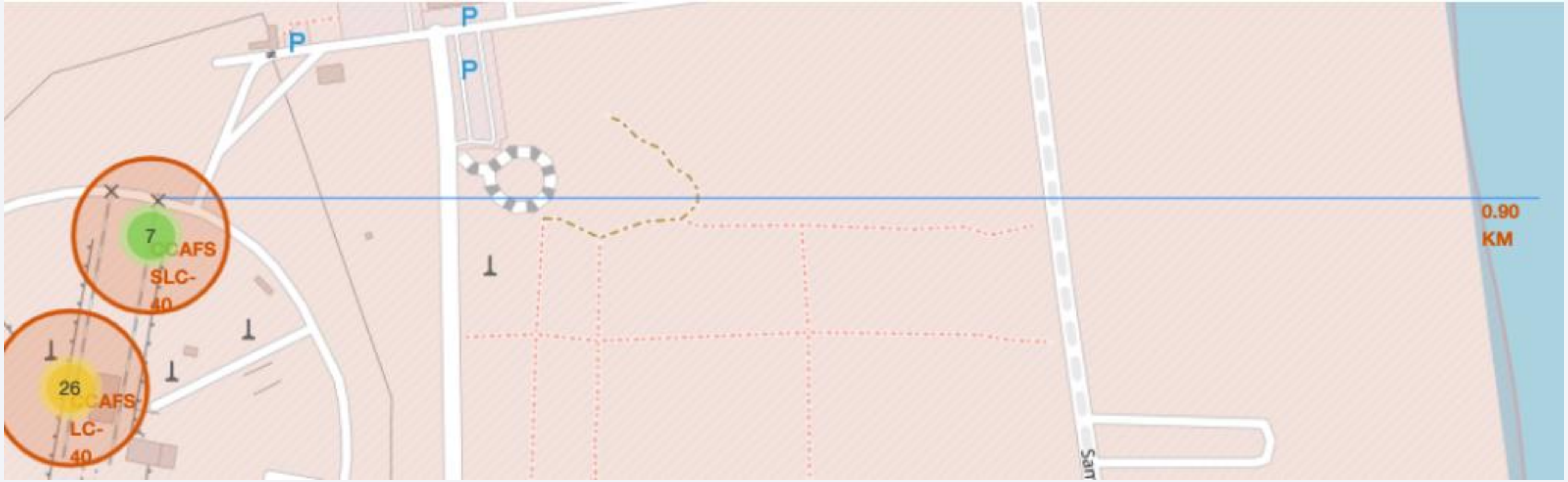
SpaceX launch sites are located near equator line and coastline - it takes advantage of the Earth's substantial rotational speed. Sitting on the launch pad near the equator, it is already moving at a speed of over 1650 km per hour relative to Earth's center.

Success/Fail Landing Outcomes - Folium

Markers on the interactive map present successful (green) and failed (red) first stage landing outcomes



Distance to the coastline - Folium



All sites are located very close the coastline, highways and railroads but far away the cities as it is potentially safer and more environmentally friendly

Section 4

Build a Dashboard with Plotly Dash

I've got an error while launching a dash app – no response from the staff yet

- Replace <Dashboard screenshot 1> title with an appropriate title

- Show the screenshot of launch success count for all sites, in a piechart

- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title

- Show the screenshot of the piechart for the launch site with highest launch success ratio

- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title

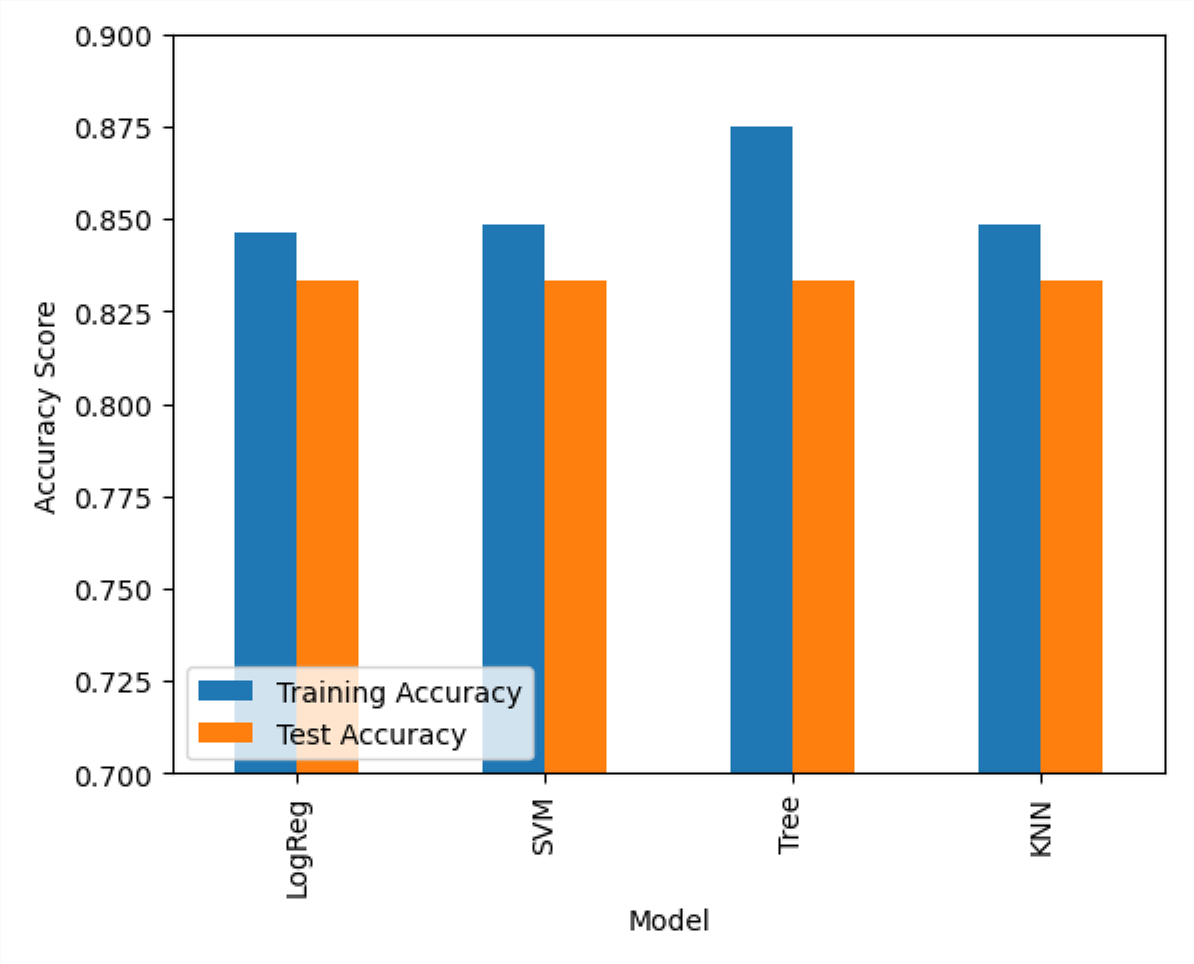
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider

- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy



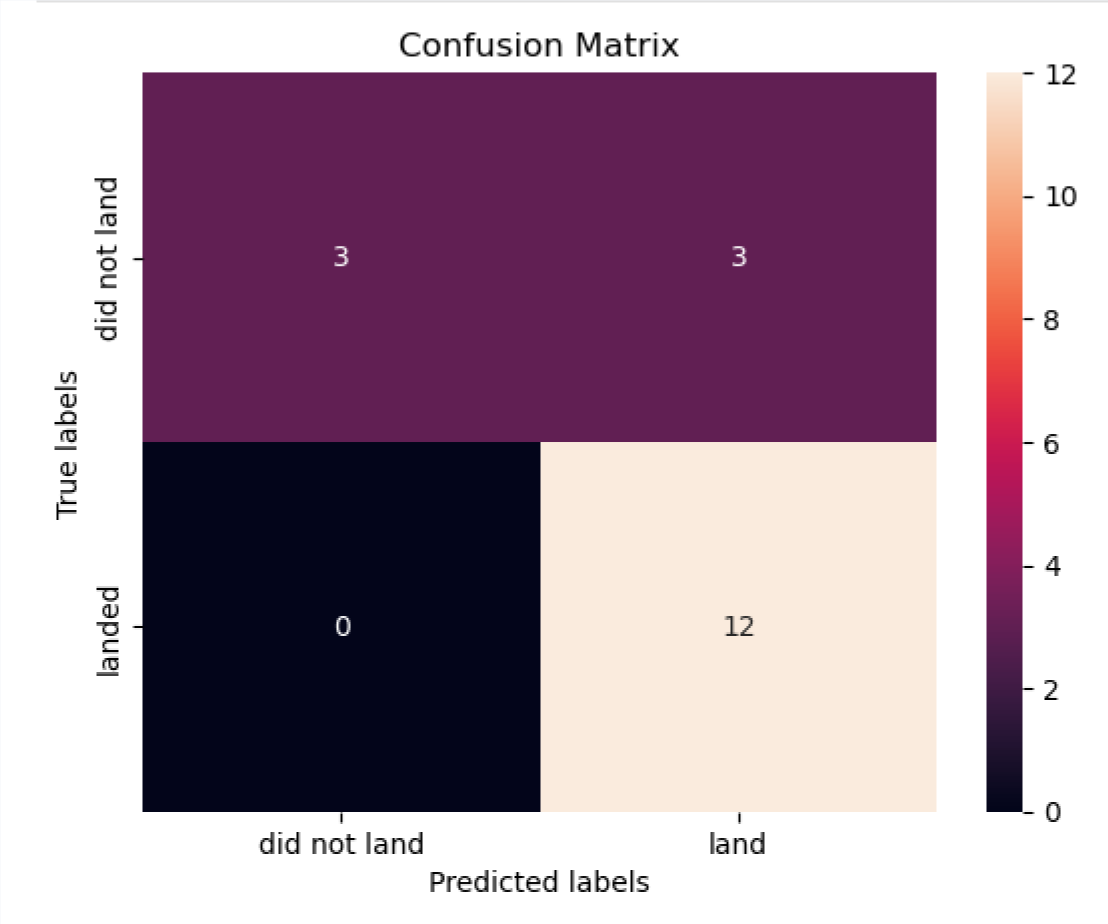
All models performed the same on the test set (0.83).

The decision tree classification model overperformed other models on the train set (0.87)

Confusion Matrix

The decision tree model correctly predicted 12 successful landings and 3 failed landings.

In 3 cases the model predicted false positive, thus predicted the first stage will land when it didn't.



Conclusions

- The overall first stage landing success for SpaceX is 0.67, however it increases with time being over 0.8 in 2020
- Landing success is highly related to payload mass, orbit, and launch site and those features interact
- Launching sites are located near equator line (as it takes advantage of the Earth's substantial rotational speed), as well as coast line but far away from cities (it is potentially safer and more environmentally friendly)
- Classification tree model developed during this project can predict if the first stage will land with 0.83 accuracy

Appendix

- Sources:
  - SpaceX APIs:
    - 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
  - Wikipedia
    - https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

Thank you!