IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Ferdi Fadillah

10/07/2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data collection with API and web scraping
  - Data wrangling to make raw data can be processed
  - Explore with EDA data visualization
  - Analyze with SQL query to answer business question
  - Data modelling with log reg, svm, knn, and decision tree
- Summary of all results
  - EDA: Launch sucess increases over time, KSC LC-39A is the highest success rate
  - Visualization: Near cost, far away from city
  - Predictive model: All model is same for testing, but for the training, decision tree is the best

Introduction

- Project background and context

SpaceX, a leader in the space industry, strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the international space station, launching a satellite constellation that provides internet access and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

- Problems you want to find answers
    - How each features affect the first-stage landing success
    - Success rate over time
    - How to predict it

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - API and web scraping
- Perform data wrangling
  - Filter, handling NaN, apply one hot encoding
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models: With train test split, GridCV, and evaluation metrics

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Request data
- json_normalize()
- Request information
- Create dictionary
- Create dataframe
- Filter dataframe
- Handle missing values
- Export

Data Collection - Scraping

- Request
- Create BeautifulSoup object
- Extract column names
- Collect data
- Create dict
- Create df
- Export

Data Wrangling

- Filter
- NaN handling
- One Hot encoding

EDA with Data Visualization

- Bar chart for standard one to know the distribute of data
- Pie Chart for persentase
- Scatter plot for relationship

EDA with SQL

- Display
- Count
- Complex query with sub query to answer business question

Build an Interactive Map with Folium

- Circle, marker, and line for the launch site

Build a Dashboard with Plotly Dash

- Dropdown to select one type of launch sites
- Pie chart to see the percentage
- Scatter plot to see the relationship

Predictive Analysis (Classification)

- Train test split
- Decide the model
- Grid CV
- Evaluate the metrics

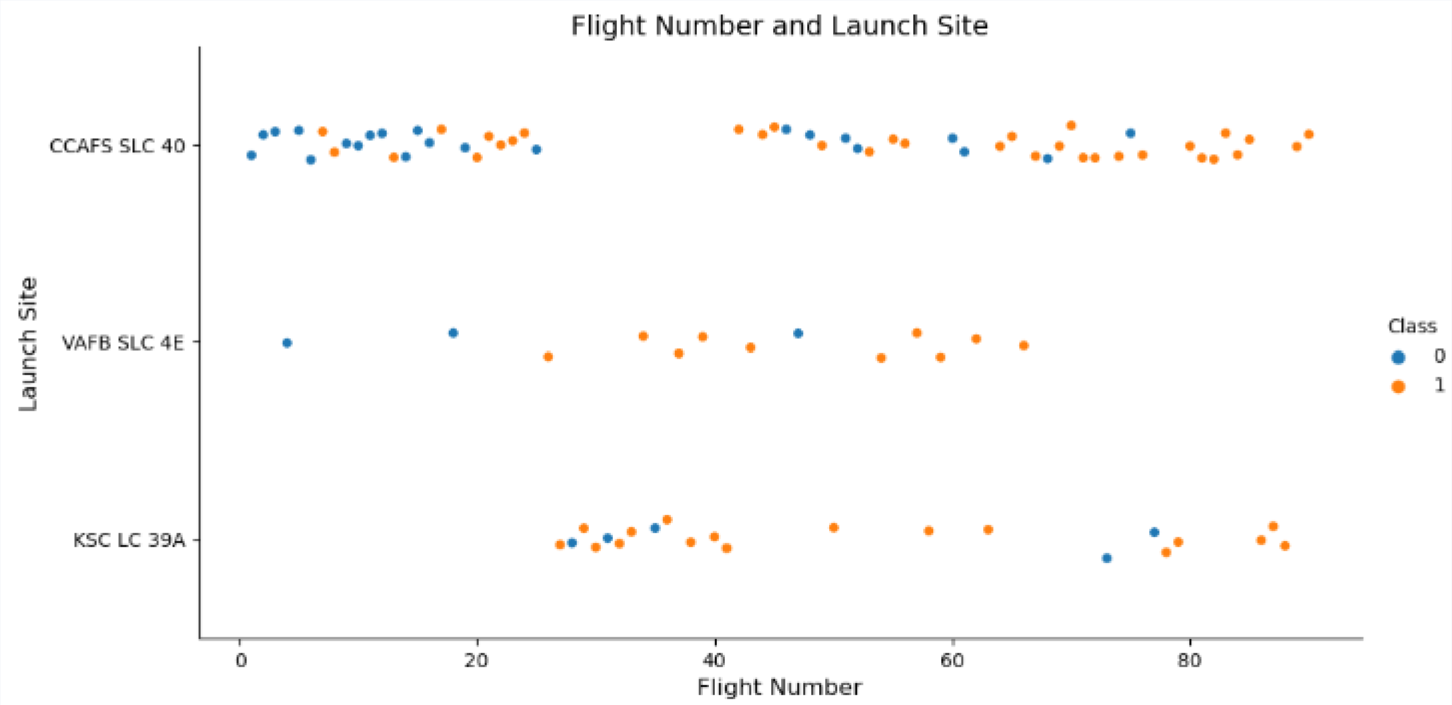
Results

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

Section 2

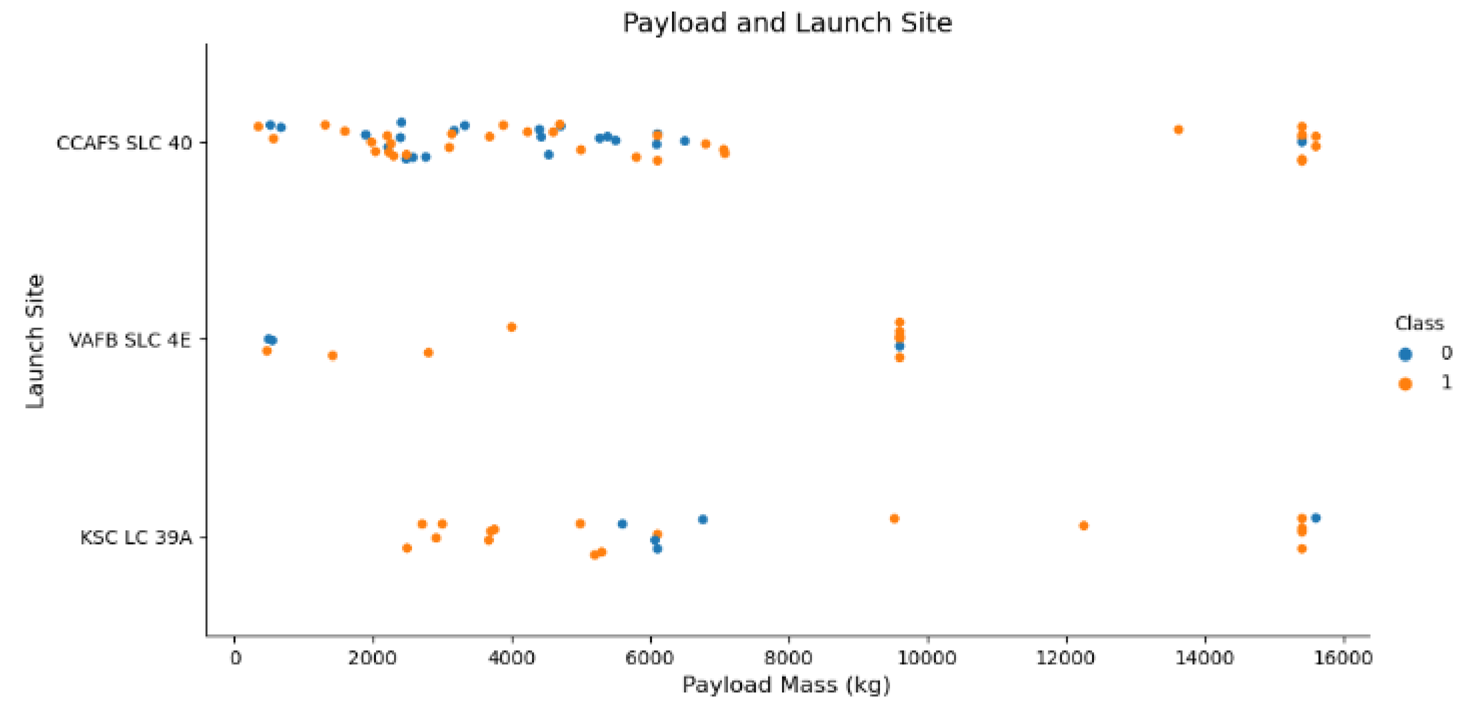
Insights drawn from EDA

Flight Number vs. Launch Site



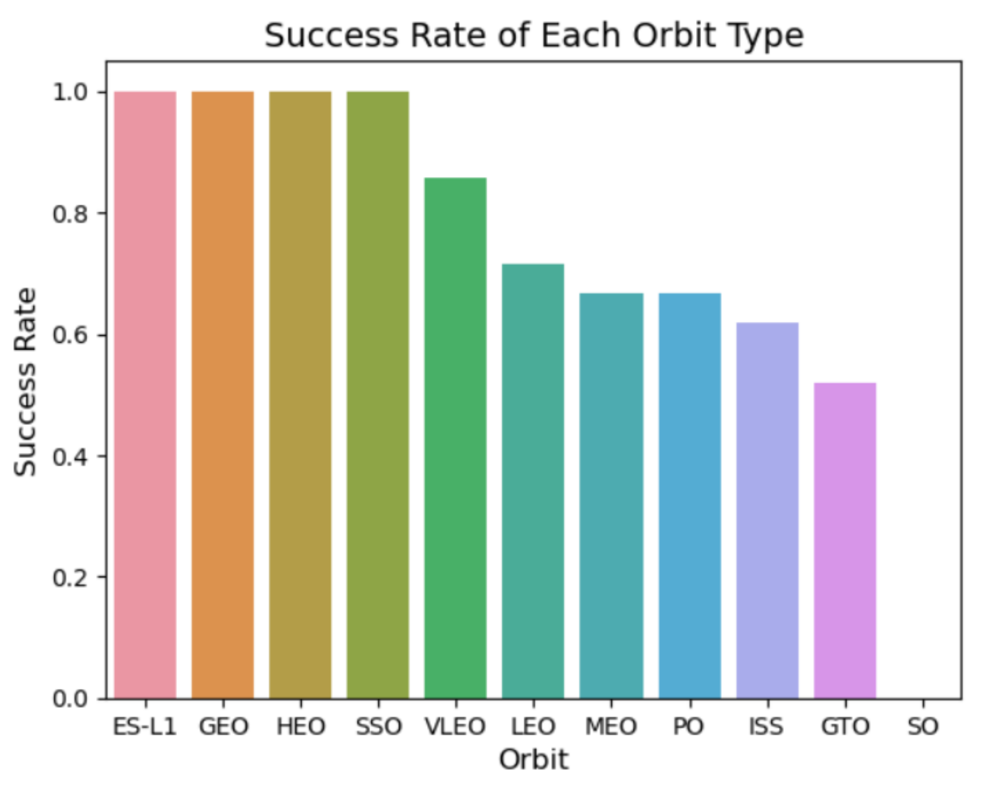
We see that different launch sites have different success rates. CCAFS LC-40, has a success rate of 60 %, while KSC LC-39A and VAFB SLC 4E has a success rate of 77%

Payload vs. Launch Site



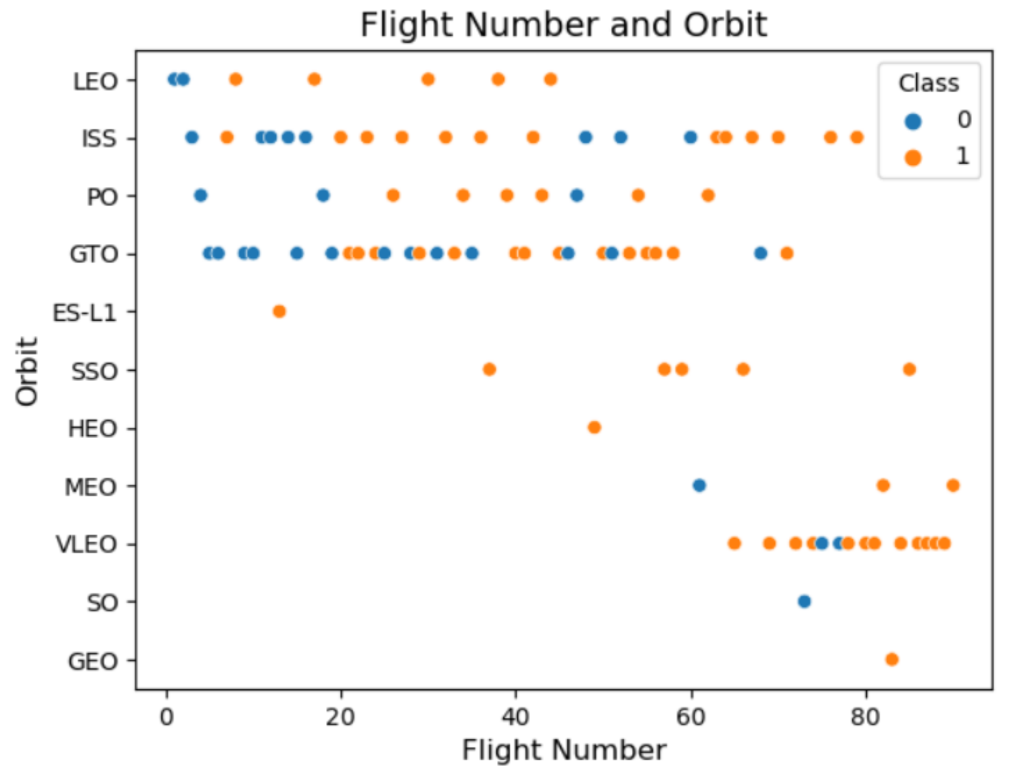
VAFB-SLC launchsite there are no rockets launched for heavypayload mass(greater than 10000)

Success Rate vs. Orbit Type



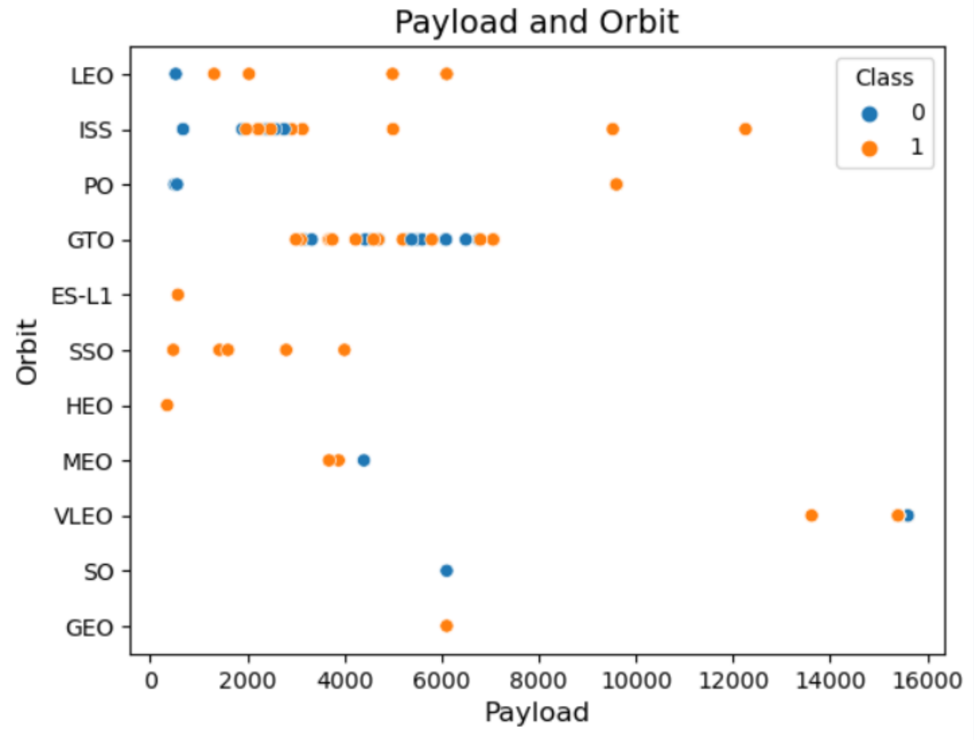
Orbits have high sucess rate: ES-L1, GEO, HEO, SSO

Flight Number vs. Orbit Type



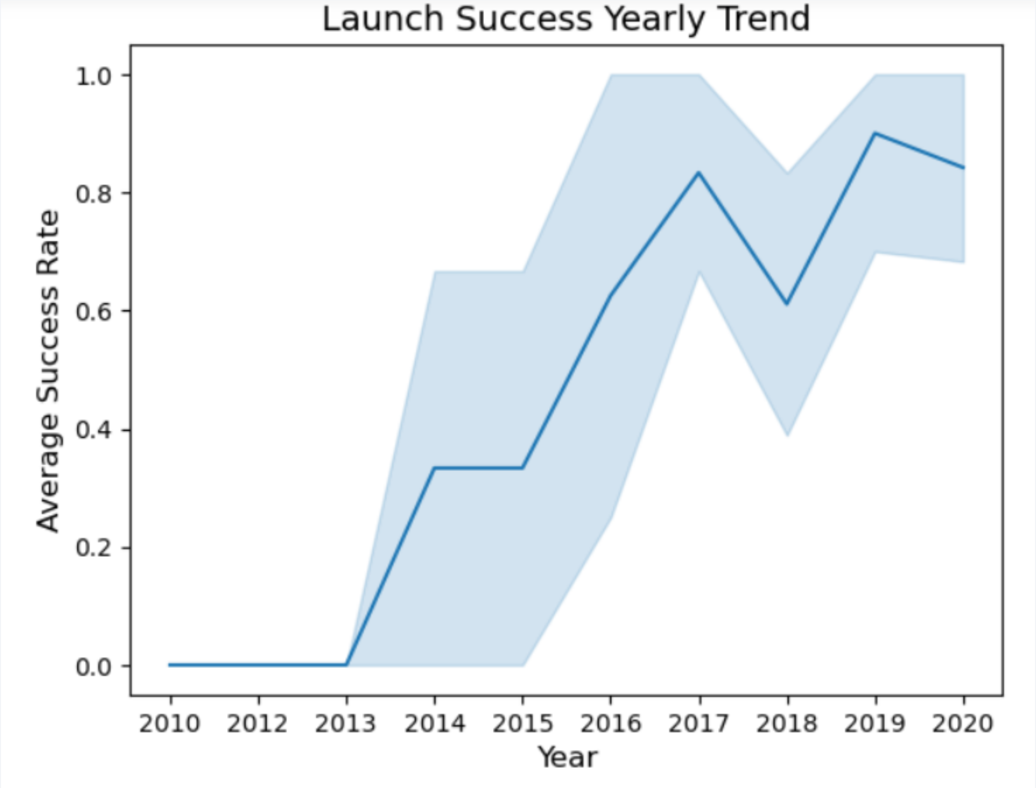
In the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit

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

There are 4 launch sites in SPACE X, those are CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)

The overview of two launch sites (CCAFS LC-40, CCAFS SLC-40)

Total Payload Mass

total payload mass
45596.0

- The total payload mass from all launch sites is 45596 kg

Average Payload Mass by F9 v1.1

average payload mass
14642.0

The average payload mass of F9 v1.1 booster version launch site is 14642

First Successful Ground Landing Date

first succesful landing outcome
01/08/2018

First successful ground landing is in 01/08/2018

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

The booster versions that are successful landed in drone ship with payload mass between 4000-6000 are F9 FT B1022, F9 FT B1026, F9 FT B1021.2, F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

```
%%sql

SELECT
    CASE
        WHEN mission_outcome LIKE '%Success%' THEN 'Success'
        WHEN mission_outcome LIKE '%Failure%' THEN 'Failure'
    END AS outcome_category,
    COUNT(*) AS count
FROM SPACEXTBL
WHERE mission_outcome LIKE '%Success%' OR mission_outcome LIKE '%Failure%'
GROUP BY outcome_category;
```

```
 * sqlite:///my_data1.db
Done.
```

outcome_category	count
Failure	1
Success	100

Only 1 failure from 101 mission outcomes

Boosters Carried Maximum Payload

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

That is the list of booster version name that carried the max payload mass.

2015 Launch Records

```
%%sql

SELECT
    CASE SUBSTR("Date", 4, 2)
        WHEN '01' THEN 'January'
        WHEN '02' THEN 'February'
        WHEN '03' THEN 'March'
        WHEN '04' THEN 'April'
        WHEN '05' THEN 'May'
        WHEN '06' THEN 'June'
        WHEN '07' THEN 'July'
        WHEN '08' THEN 'August'
        WHEN '09' THEN 'September'
        WHEN '10' THEN 'October'
        WHEN '11' THEN 'November'
        WHEN '12' THEN 'December'
    END AS monthnames,
    Landing_Outcome,
    Booster_Version,
    Launch_Site
FROM SPACEXTBL
WHERE SUBSTR("Date", 7, 4) = '2015' AND Landing_Outcome = 'Failure (drone sh
```

```
 * sqlite:///my_data1.db
Done.
```

monthnames	Landing_Outcome	Booster_Version	Launch_Site
October	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

The failure launch records in 2015

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

Landing_Outcome	count of landing outcomes
Success (ground pad)	7
Failure (drone ship)	3

The rank of landing outcomes (success and failure) between aforementioned period

Section 3

Launch Sites Proximities Analysis

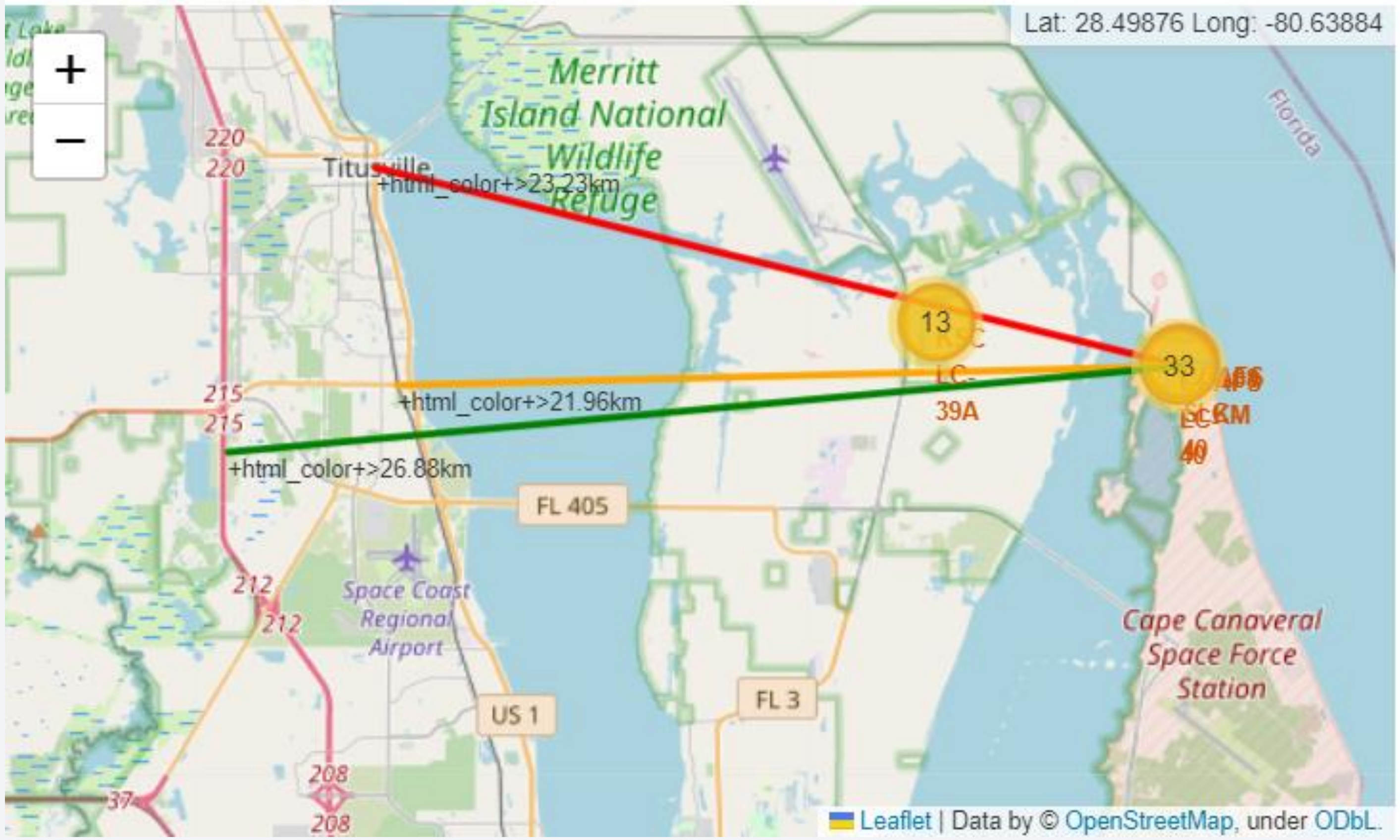
All launch sites on a map

All launch sites are marked with red circle and title

Success/failed launches for each site on the map

For each site on the map, we can see the overview of success/failed launches with red and green icon

Distances between a launch site to its proximities

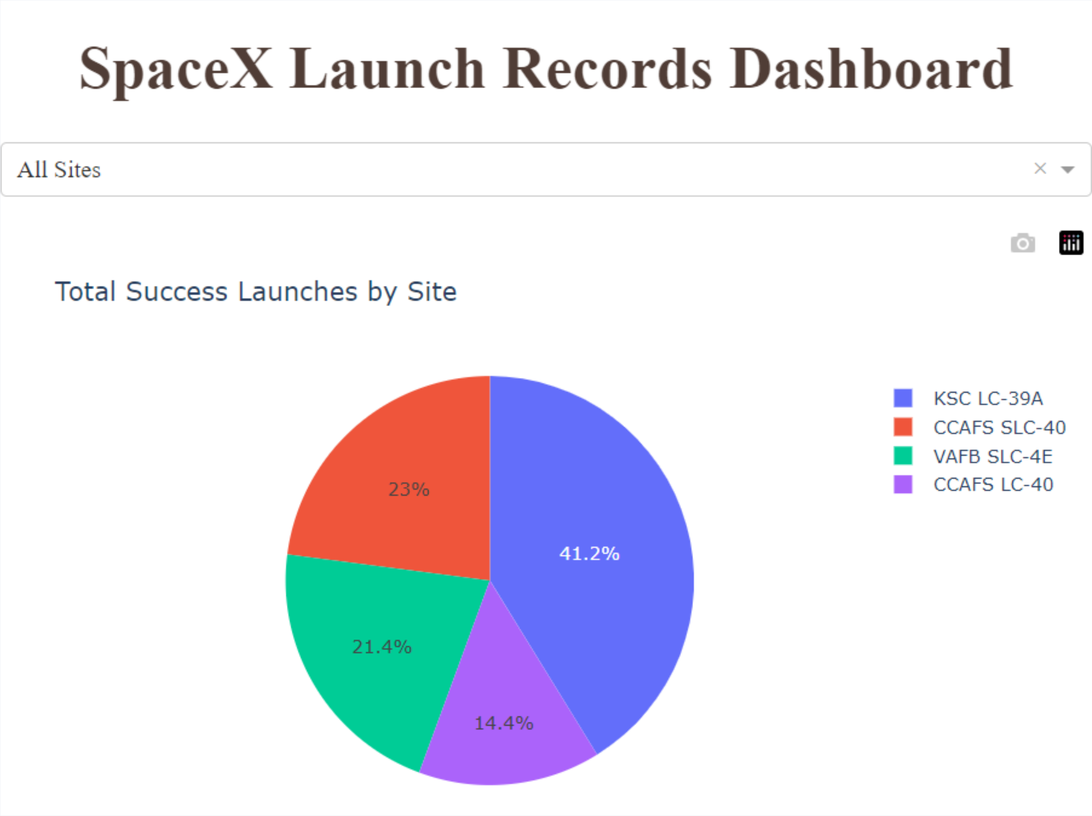


Launch site distance to city, railway, highway, and coastline (see the colored line)

Section 4

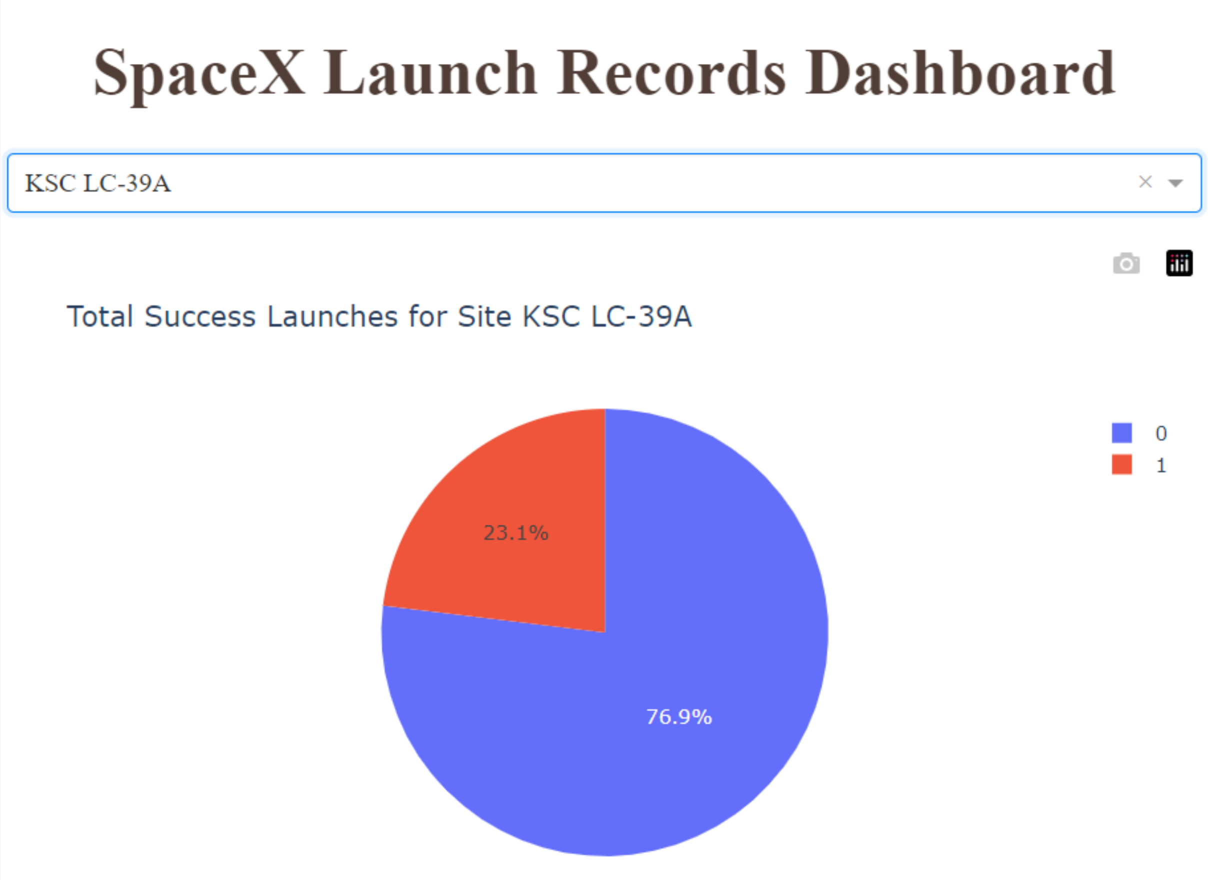
Build a Dashboard with Plotly Dash

Launch success count for all sites



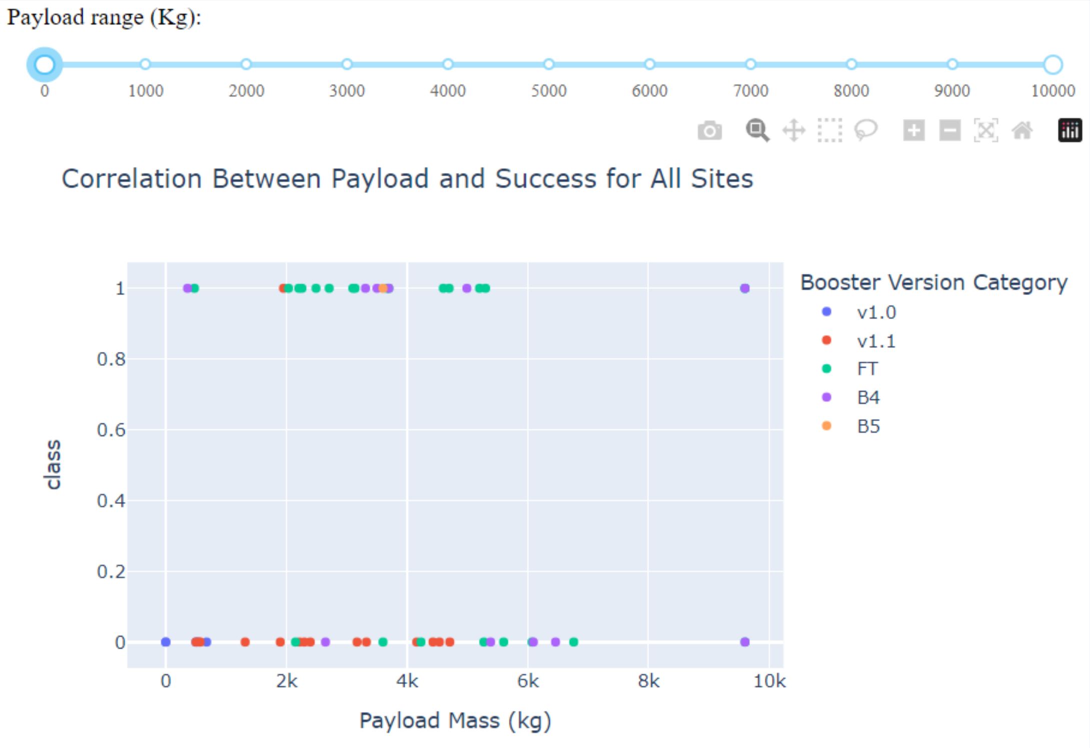
The most of success launch is dominated by KSC LC-39A for 41.2%.

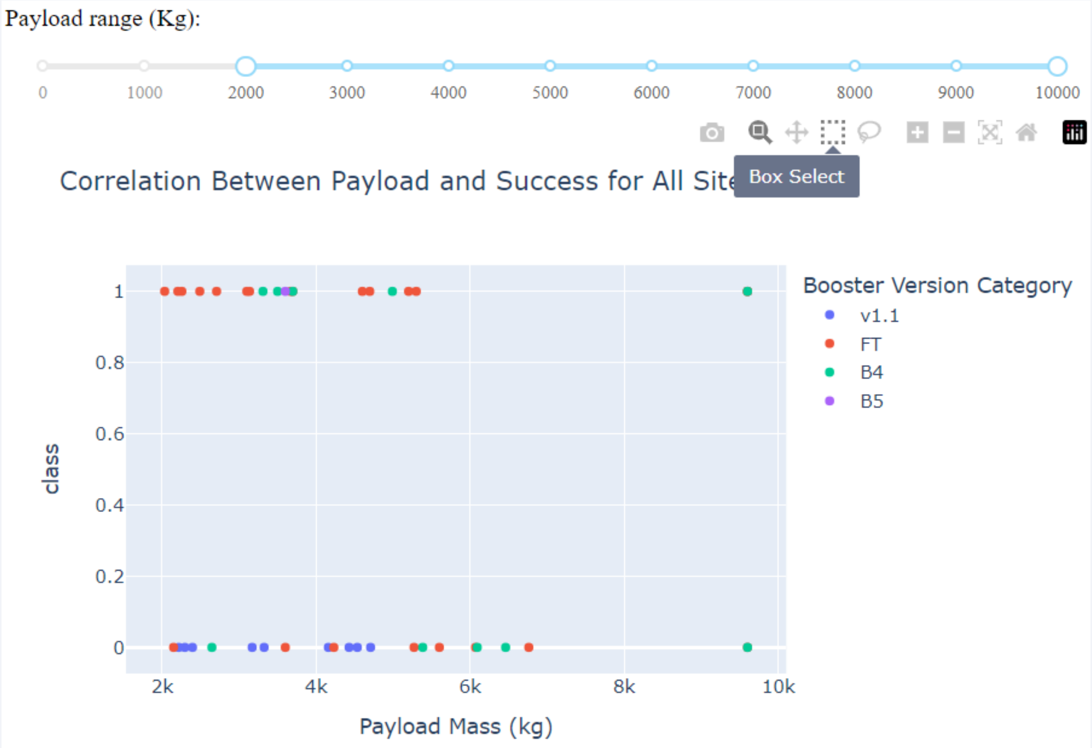
Launch site with highest launch success ratio



The highest launch success rasio is KSC LC-39A for 76,9%

Payload vs. Launch Outcome scatter plot for all sites



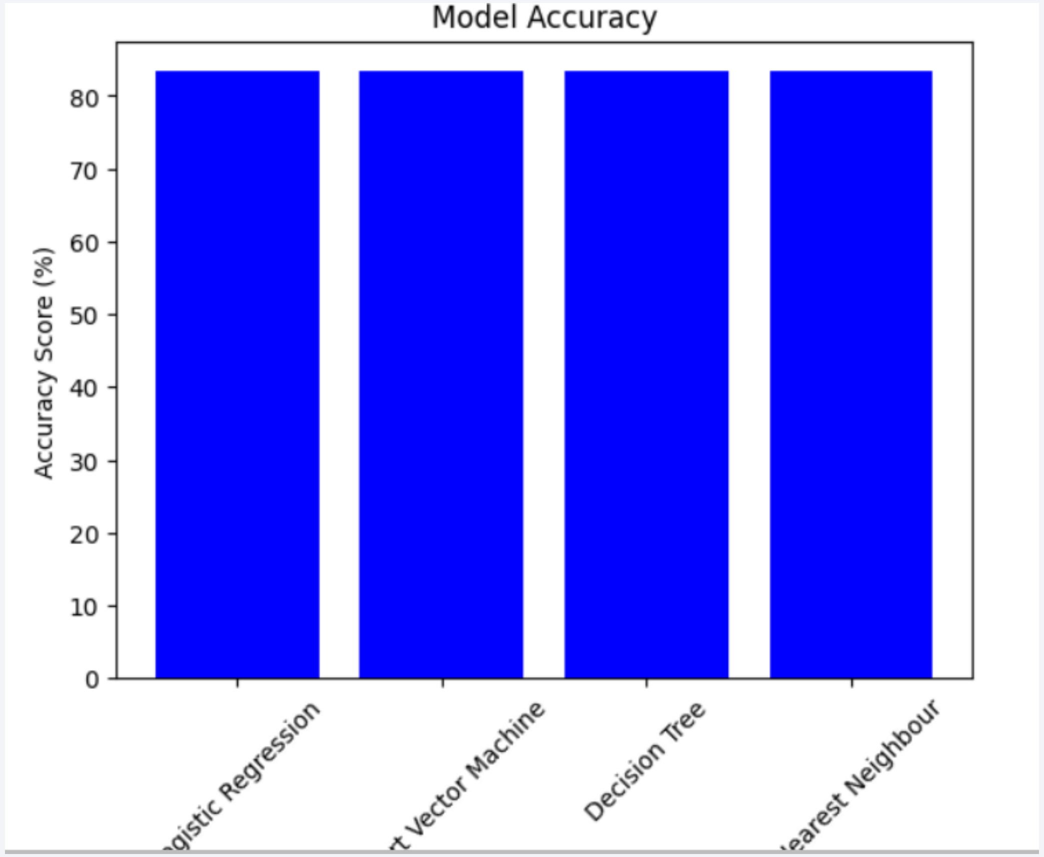


For each payload range slider, the majority of booster version category is also different. In example, 0 slider contains mostly B4 and 2000 slider contains mostly FT. For the specific numeric, we can calculate scratchly instead of visualization

Section 5

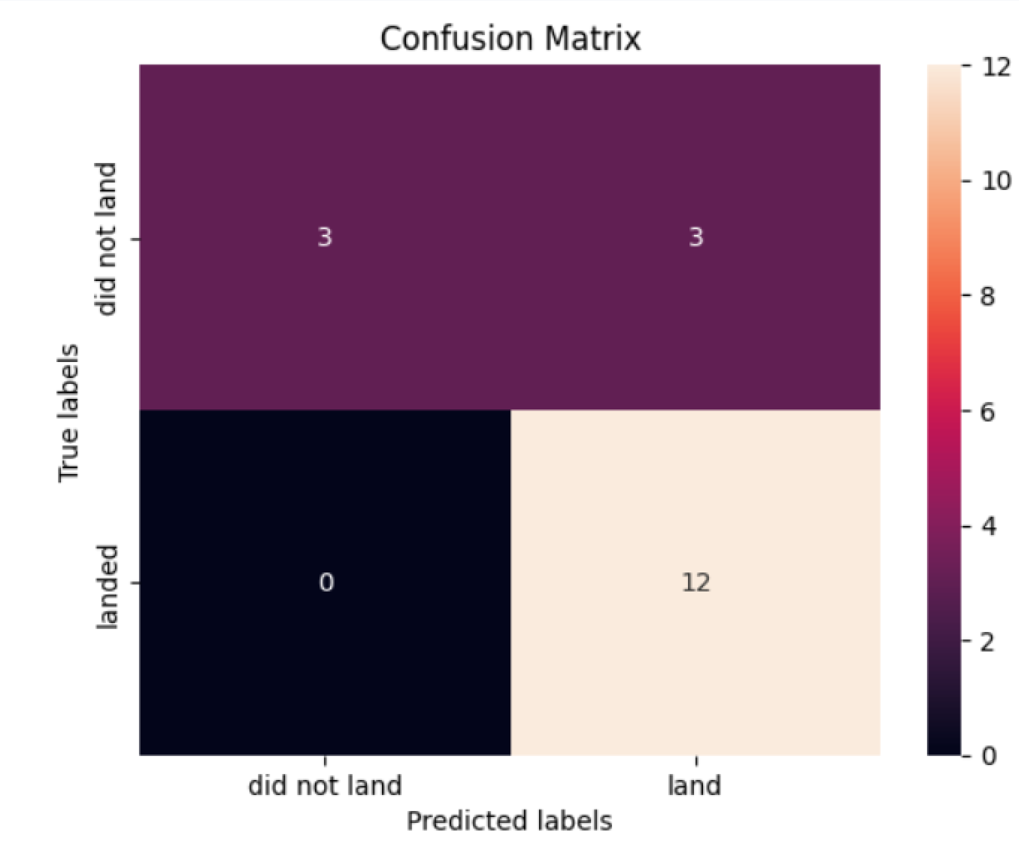
Predictive Analysis (Classification)

Classification Accuracy



All are same for testing because the data is small

Confusion Matrix



The confusion matrix is unbalance, maybe the data is unbalance and too small

Conclusions

- Launch sites is near to the cost and far away from city, highway, and railway for safety reason
- KSC LC-39A has the highest success rate
- The sucess rate since 2013 kept increasing till 2020
- Orbits ES-L1, GEO, HEO, and SSO have 100% success rate
- The higher payload mss, the higher the success rate
- Each model has same accuracy because small dataset
- Or, we can do feature engineering
- Or, we can choose another model

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Github link:

https://github.com/ferdifdi/ibm-data-science/tree/main/applied_data_science_capstone

Thank you!