IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

AKHIL SAI MALLAMPATI
21/03/2023

Outline

EXECUTIVE SUMMARY

INTRODUCTION

METHODOLOGY

RESULTS

CONCLUSION

APPENDIX

Executive Summary

Summary of methodologies

Data Collection & Data Wrangling

EDA w/ Data Visualization & SQL

Interactive Map w/ Folium

Dashboard w/ Plotly Dash

Classification Analysis

Summary of all results

Exploratory data analysis results

Interactive analytics maps and dashboards

Predictive analysis results

Introduction

Project background and context

The project goal is to predict the likeliness of Falcon 9's landing success rate. SpaceX promotes Falcon 9 rocket launches on its website, costing 62 million dollars; other providers cost upwards of 165 million dollars each, and 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

Problems you want to find answers

- What are the criteria for a successful first-stage landing?
- What are the different variables impacting the success rate of a successful landing?
- What conditions does SpaceX have to accomplish to get the optimal rocket success landing rate?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Utilize SpaceX Rest API
  - Wikipedia Web Scraping
- Perform data wrangling
  - Clean unrelated & dirty data
  - Transform data through One Hot Encoding for Machine Learning
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Plotting – Scatter, Bar, and line graphs to show patterns of data
- Perform interactive visual analytics using Folium and Plotly Dash
  - Create interactive maps and dashboards for visual purposes
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data sets were collected using the SpaceX Rest API

The API data regarding rocket launches included information such as:

- Rocket Type
- Payload Delivery
- Launch/Landing Specification
- Landing Outcome

Wikipedia Web Scraping

Data Collection - SpaceX API

Response from API

Convert Response to .json file

Apply functions to clean data

Assign list to dictionary then data frame

Filter data frame and export to flat file

SpaceX Rest API call returned data in JSON format

Data that was returned was passed on to the dictionary

Dictionary then was converted to a Pandas data frame

Once converted, it was exported as CSV file

https://github.com/AKHILSAIM/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20API%20Lab.ipynb

Data Collection - Scraping

Getting response from HTML

Creation of BeautifulSoup Object

Extract data into dictionary

Convert dictionary to data frame

Export data frame to CSV file

- Utilized Wikipedia to scrape data
- URL is requested to contain data
- HTML response is written to pass into BeautifulSoup
- Extract data into the dictionary
- It is then converted into Panda data frame
- Data frame is exported to CSV

https://github.com/AKHILSAIM/Applied-Data-Science-Capstone/blob/main/Data%20Collection%20with%20Web%20Scraping%20lab.ipynb

Data Wrangling

Exploratory Data Analysis

Identifying Training Target Feature

Assigning labels to the Target Feature

Calculate the count of launches per site and orbit

Create a landing outcome label from Outcome column

Through the data wrangling process, you can dictate the training labels from the raw data

EDA steps were executed by examining the percentage of categorical and numerical values

Once EDA was accomplished, the training labels were assigned to the Landing Outcome Feature

https://github.com/AKHILSAIM/Applied-Data-Science-Capstone/blob/main/Data%20Wrangling%20with%20EDA%20Lab.ipynb

EDA with Data Visualization

Scatter Plots

Convey correlation between features such as:

- Flight number, payload mass, flight number, launch site, payload/launch site, etc.

Bar Graph

Compare data sets between different categories such as:

- Landing success for each orbit type

Line Graph

Show trends to clearly mark the tendency of the data and make a prediction

- Average success rate per year

https://github.com/AKHILSAIM/Applied-Data-Science-Capstone/blob/main/EDA%20with%20Visualization%20la

EDA with SQL

- SQL was performed to collect information from the dataset by requesting 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 average payload mass carried by boosters launched by NASA
  - Display average payload mass carried by booster version F(v1.1
  - List the date when the first successful landing outcome in the ground pad was achieved
  - List the names of the boosters which had success in drone ship
  - List the total number of successful and failure mission outcomes
  - List the names of the booster versions which have carried the max payload mass
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Build an Interactive Map with Folium

Utilized the latitude and longitude coordinates for each launch site and added a circle marker around each launch site with a label of the name of the launch site

Markers were color-coded with the launch outcome labels

Green marker was used to determine if a launch was successful and the red marker when a launch was failed

- Green marker for 0
- Red marker for 1

The Haversine's formula was used to calculate the distance from the Launch site to the marked landmarks

https://github.com/AKHILSAIM/Applied-Data-Science-Capstone/blob/main/Interactive%20Visual%20Analytics%20with%20Folium%20lab.ipynb

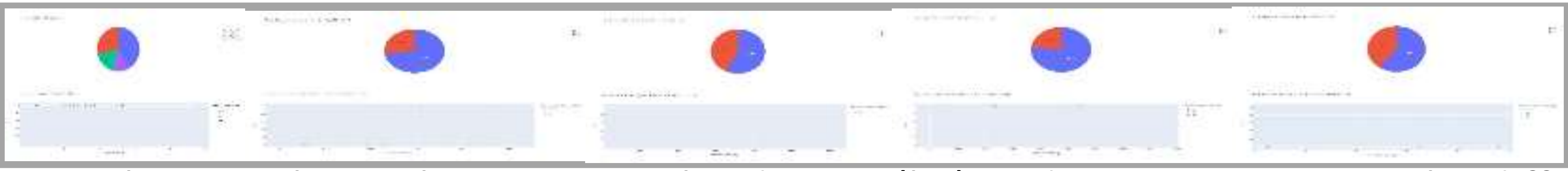
Build a Dashboard with Plotly Dash

For the project, both the pie chart and scatter plot was created to demonstrate the success rate of the landing

Pie Chart:

- Shows the proportion between success and failures of launch site by type

Scatter Plot:



- Graphs portray the correlation between payload mass (kg) and mission outcome by different categories such as sites and booster version

https://github.com/AKHILSAIM/Dashboard-Application-with-Plotly-Dash.git

Predictive Analysis (Classification)

- 3 Steps process for the predictive analysis procedure:
  - Building model:
    - Load dataset into pandas
    - Transform data
    - Fit datasets into GridSearchCV objects and train dataset
  - Evaluating Model
    - Check the accuracy for each dataset and model created
    - Plot Confusion Matrix
  - Finding the best-performing classification model
    - Decide which model set has the best accuracy and choose

https://github.com/AKHILSAIM/Applied-Data-Science-Capstone/blob/main/Machine%20Learning%20Prediction%20lab.ipynb

Results

EXPLORATORY DATA ANALYSIS RESULTS

INTERACTIVE ANALYTICS DEMO IN SCREENSHOTS

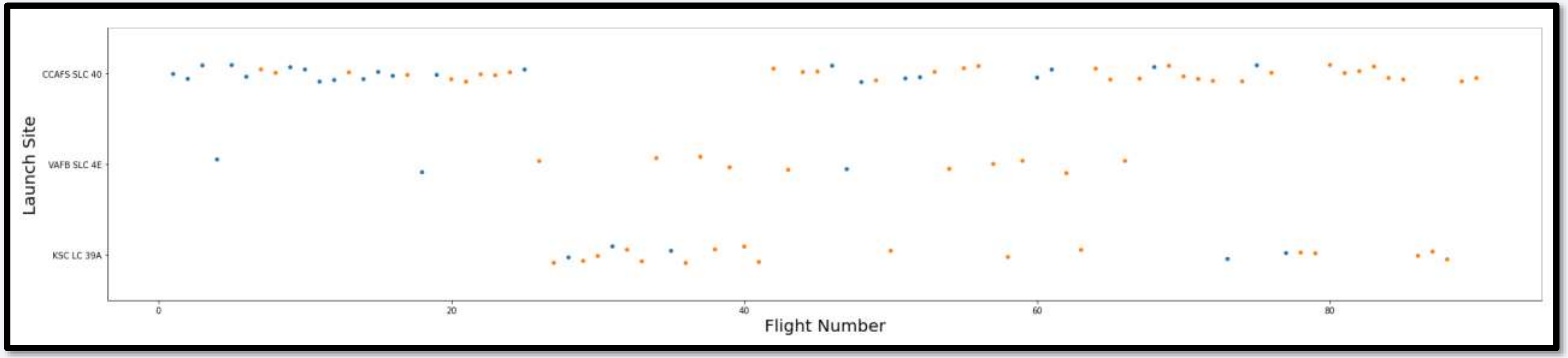
PREDICTIVE ANALYSIS RESULTS

Github URL

Section 2

Insights drawn from EDA

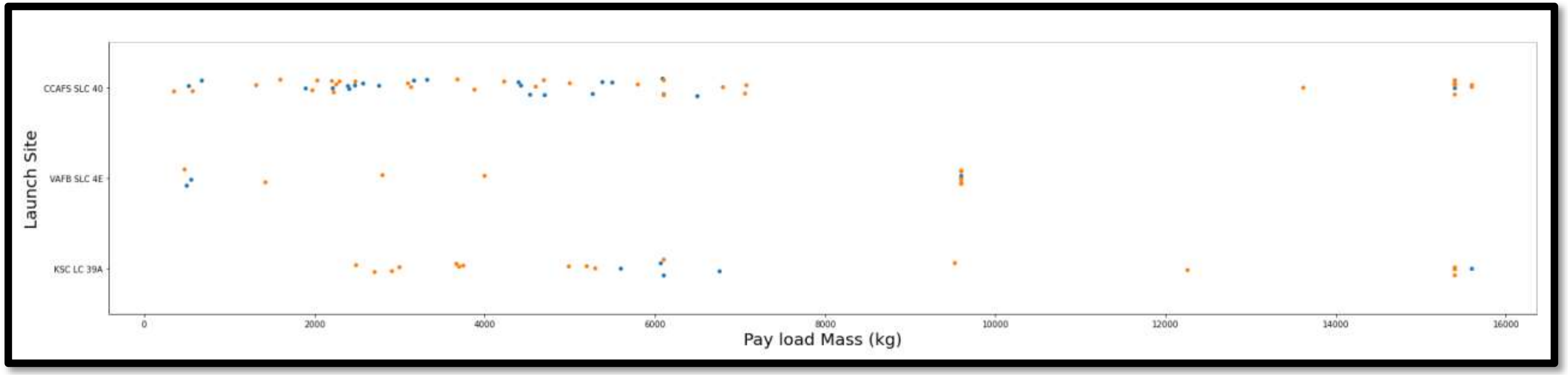
Flight Number vs. Launch Site



- Recent missions have proven to be more successful at every launch sites

Payload vs. Launch Site

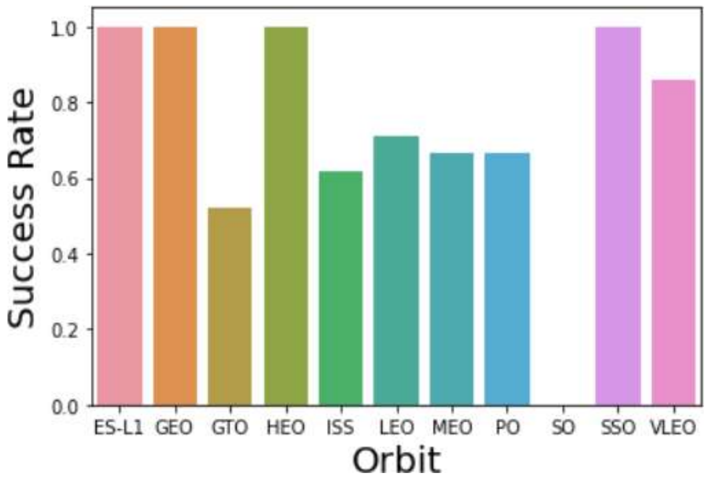
- There is a correlation between payload mass and the success rate
  - As payload mass increases, the success rate will be higher as well

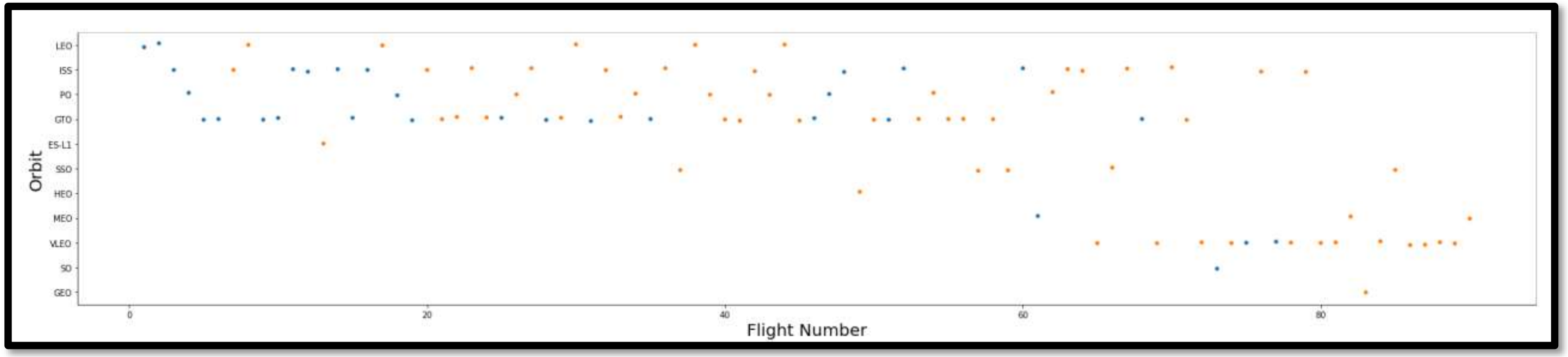


Success Rate vs. Orbit Type

- Stages from the first round landed successfully from ES-L1, GEO, HEO, SSO orbits

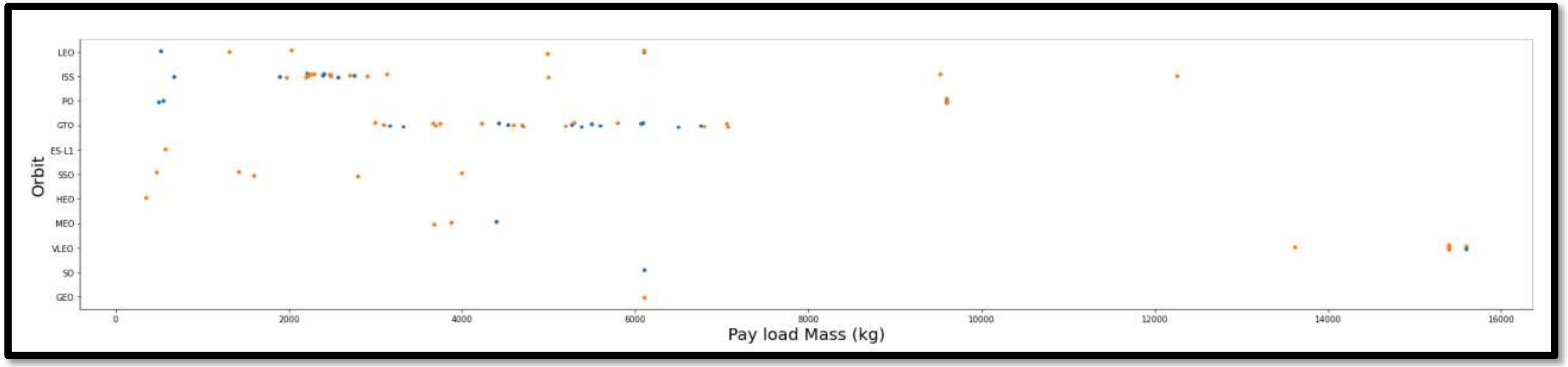
- The success rate at SO orbit is zero





- The GTO Orbit does not have any connection with flight number
- VLEO orbit has a high success rate
- Majority of high flight numbers go to GEO, SO, VLEO, MEO orbits

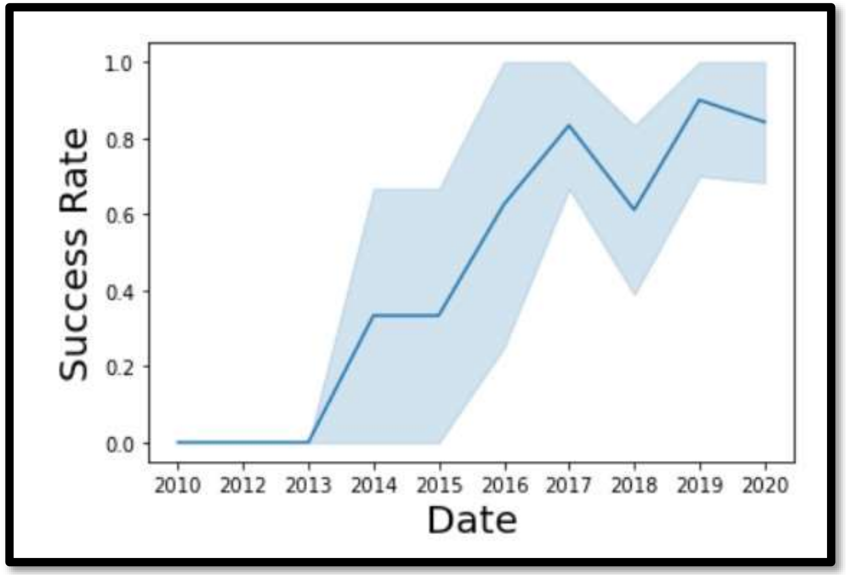
Flight Number vs. Orbit Type



- SSO orbit has a higher success rate when payload mass is lower
- GTO orbit has no connection with the payload mass
- The LEO orbit has a higher success rate when payload mass is higher

Payload vs. Orbit Type

Launch Success Yearly Trend



- There were no successful landings before 2013
- The success rate continues to jump after 2013

All Launch Site Names

- SQL Query
  - Select distinct launch_site from spacextbl

- Clarification:
  - Purpose is to search for names of unique launch sites in the space mission
  - The distinct statement will show the unique launch site names as unique values

Query Result

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

Launch Site Names Begin with 'CCA'

- SQL Statement used:
  - Select * from spacextbl where launch_site like ‘CCA’ limit 5
- Explanation:
  - Searching for 5 records where launch sites begin with CCA
  - Used like operator to search for keywords in launch site column
  - Used limit clause to show only 5 records
- Query Result:

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

SQL Statement:

```
Select sum (payload_mass_kg_)
total_payload_mass from spacextbl where
customer = 'NASA (CRS)'
```

Explanation:

The query is searching for the total payload mass carried by boosters launched by NASA (CRS)

Sum function was used to add all the payload mass

Query Result:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

Select Statement:

Select avg (payload_mass_kg_) avg_payload_mass
from spacextbl where booster_version = ‘F9 v1.1’

Explanation:

With the query, we are searching for the average payload mass carried by F9 v1.1

Used the average function to calculate the average payload mass

Query Result:

avg_payload_mass
2928

First Successful Ground Landing Date

SQL Statement:

Select min(date) min_date from spacextbl where landing_outcome = 'Success (ground pad)'

Explanation:

The query is designed to search for the date when the first successful ground landing was achieved

Used the min function to return the smallest value of the date

Query Result:

min_date
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- SQL Statement:
  - Select booster_version from spacextbl where landing_outcome = 'Success (drone ship)' and payload_mass_kg_ > 4000 and payload_mass_kg_ < 6000
- Explanation:
  - The query is searching for the names of boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
  - Using the condition landing_outcome = "Success (drone ship)' and payload_mass_kg_ > 4000 and payload_mass_kg_ < 6000 to filter the action
- Query Result:

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

Total Number of Successful and Failure Mission Outcomes

mission_outcomes	qty
Failure	1
Success	100

- SQL Statement:
  - Select (case when mission outcome like ‘%Success%’ then ‘Success’ else ‘Failure’ end) mission_outcomes, count(*) qty from sapcextbl group by (case when mission outcome like ‘%Success%’ then ‘Success’ else ‘Failure’ end)
- Explanation
  - The query is designed to search for the total number of successful and failure mission outcomes
  - Used count(*) statement to retrieve the number of records
  - Used group by statement to group records by mission outcomes
- Query Results:

Boosters Carried Maximum Payload

- SQL Statement:
  - Select booster_version from spacextbl where payload_mass_kg_ = (select max(payload_mass_kg_) from spacextbl
- Explanation:
  - The query is searching for the names of booster version which have carried the max payload mass
  - Used max function to return the largest value of payload mass
- Query Result:

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

SQL Statement:

Select booster_version, launch_site from spacextbl where landing_outcome = 'Failure (drone ship)' and year(date) = 2015

Explanation:

The query is searching for the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015

Query Result:

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

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

- SQL Statement:
  - Select landing_outcome, count(landing_outcome) qty from spacextbl where (date between ‘2010-06-04’ and ‘2017-03-20’) group by landing_outcome order by 2 desc
- Explanation:
  - The query is designed to rank the count of landing outcomes between the date 6/4/2010 and 3/20/2017 in descending order
  - Used group by statement to group records by landing outcomes and the desc keyword to sort in descending order
- Query Result:

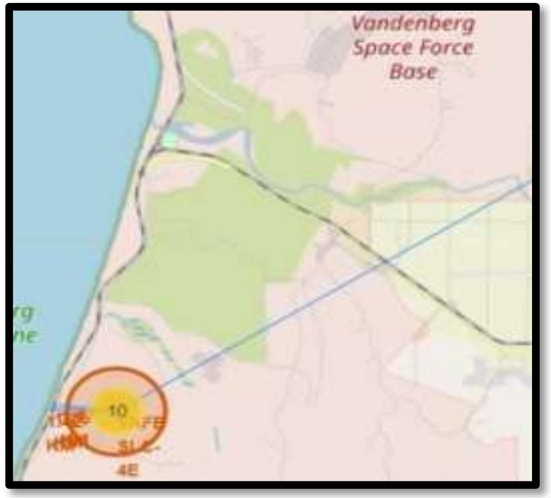
landing__outcome	qty
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

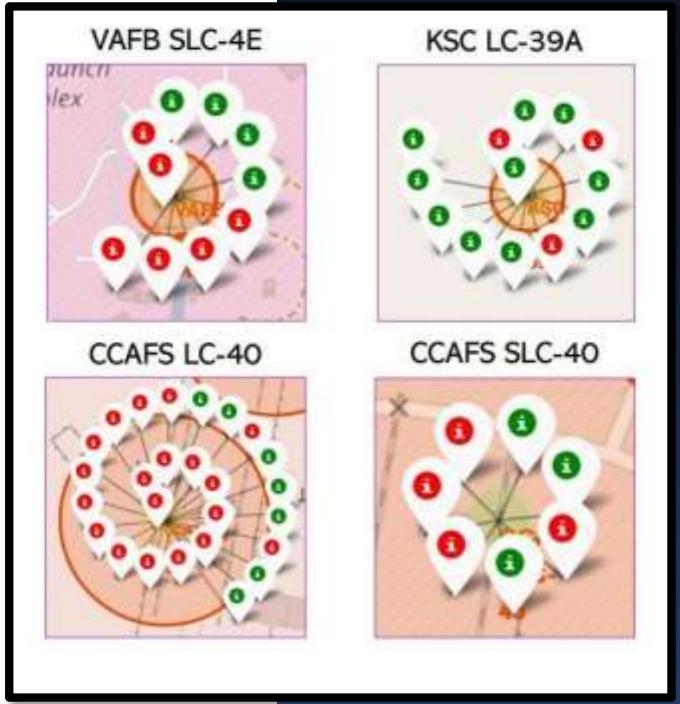
Folium Map – SpaceX Sites

- Launch sites are in proximity to the equator and all launch sites are in close proximity to the coast
- This was done in order to minimize the risks in the event of an accident
- Launch sites are in close proximity to railways, highways, and coastlines
- They are away from a distance from congested cities



Folium Map – SpaceX Launch Site Success Rate

- Markers for each landing outcome by launch site are present
- Green mark: Success
- Red mark: Failure

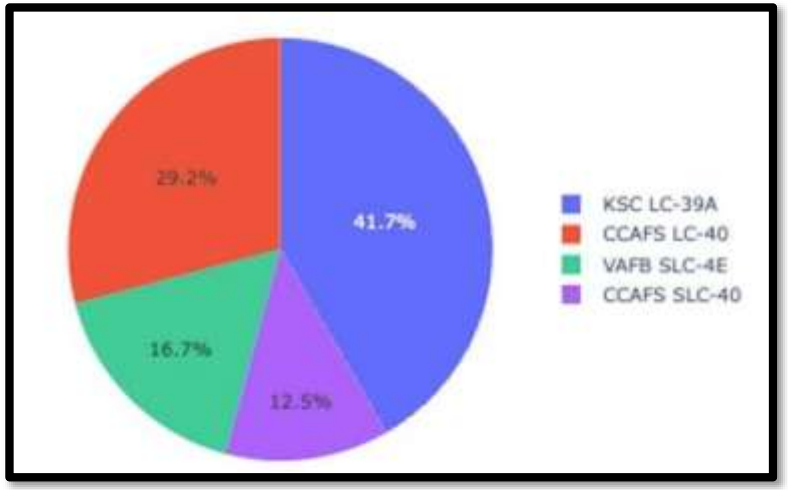


Section 4

Build a Dashboard with Plotly Dash

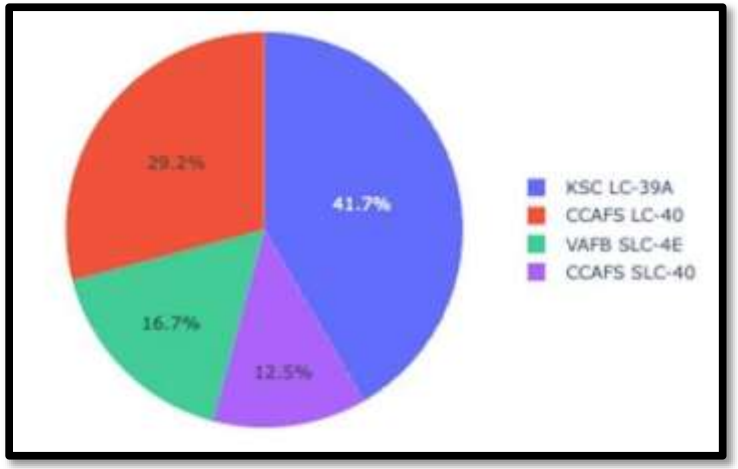
Dashboard – Launch Success Count by Sites

- Pie Chart Display
- Shows the total success of launches for all sites
- KSC LC-39A has the most successful launches



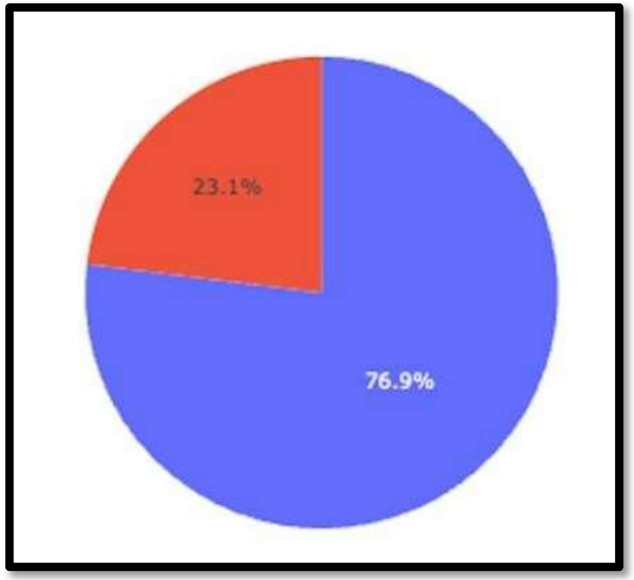
Dashboard – Launch Success Count by Sites

- Pie Chart Display
- Conveys the total success launches for all sites
- KSC LC-39A has the most successful launches



Dashboard – Launch Site w/ Highest Launch Ratio

- Pie Chart Display
- Shows the ratio of successful and failed launches by launch site
- KSC LC-39A has the highest success rate our of all sites

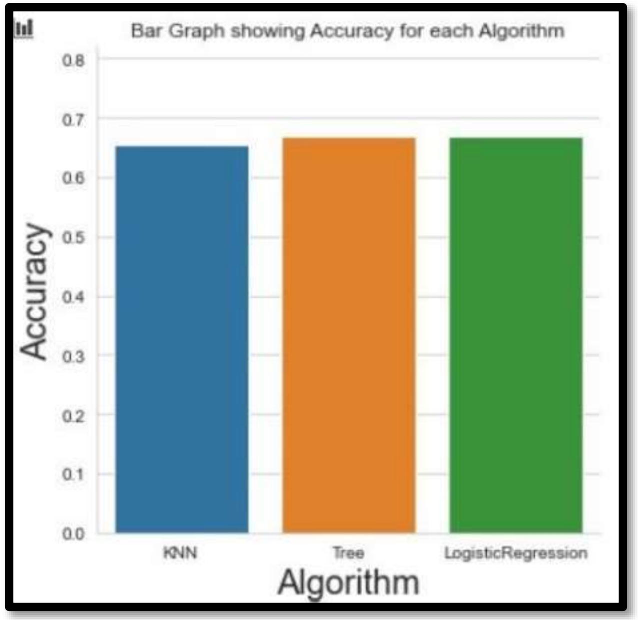


Section 5

Predictive Analysis (Classification)

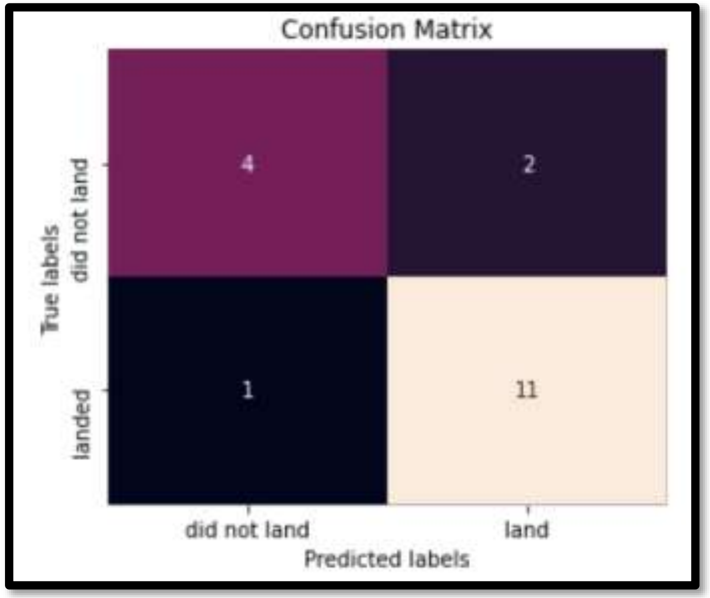
Classification Accuracy

- Classifiers' accuracy was relative. The decision Tree is found to be at peak performance with 0.8623 score
- At the end of things, the decision tree classifier achieved 83.33% accuracy on the test data



Confusion Matrix

- The decision tree model has the best balance between matches and false negatives / false positives
- Major problem is false true negative and false positive



Conclusions

The Decision Tree Model is the best for Machine Learning for this project as the model performs at an optimal level compared to others

Low-weighted payloads perform statistically better compared to heavier payloads according to the data analytics

Out of all the sites, KSC LC-39A was the most successful sites by successful launches

Out of all orbits, GEO, HEO, SSO, ES-L1 have the highest success rate per data analytics

Thank you!