IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Juan Pablo Ravetti
17/02/2024

Outline

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

Executive Summary

Summary of methodologies

- Data Collection, Web Scraping
- Data Wrangling
- Exploratory Data Analysis with Data Visualization
- Exploratory Data Analysis with SQL
- Interactive map with Folium
- Dashboard with Plotly Dash
- Predictive Analysis

Summary of all results

- Exploratory Data Analysis results
- Interactive Analysis based on maps and dashboards
- Predictive Analysis results

Introduction

Project Background and Context

SpaceX (Space Exploration Technologies Corp.) stands as the preeminent leader in the era of commercial space exploration, pioneering affordable space travel.The company 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. Based on public information and machine learning models, we are going to predict if SpaceX will reuse the first stage.

Problems you want to find answers

- What are the primary attributes defining a successful or unsuccessful landing?
- How do different rocket variables impact the outcome of a landing, whether success or failure?
- What conditions need to be met for SpaceX to attain the highest landing success rate?

Section 1

Methodology

Methodology

Executive Summary

- **Data collection methodology:**
  - SpaceX REST API
  - Web Scraping from Wikipedia
- **Perform data wrangling**
  - Dealing with null values and filtering data
  - One Hot Encoding for classification models
- **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

Data Collection

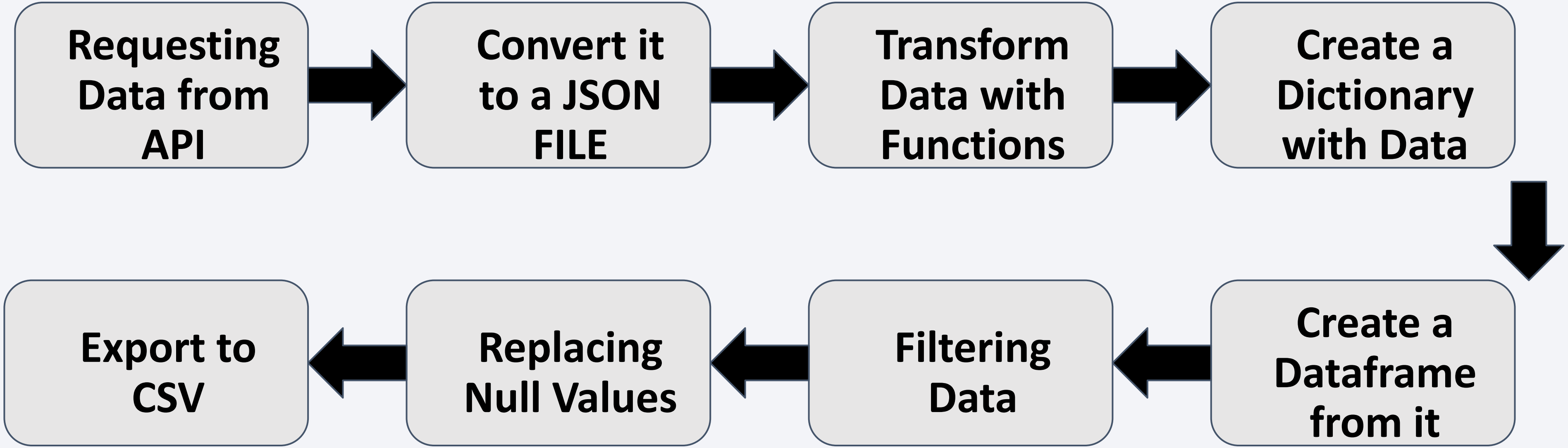
Our data collection process involved a dual approach: We utilized SpaceX's REST API for retrieving specific information and also performed web scraping to extract data from a table within SpaceX's Wikipedia entry.

This combined method ensured that we captured comprehensive information about the launches, facilitating a more thorough analysis.

- **Data Columns obtained by using SpaceX REST API:**
  FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

- **Data Columns obtained by using Wikipedia Web Scraping:**
  Flight No., Launch site, Payload, Payload mass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

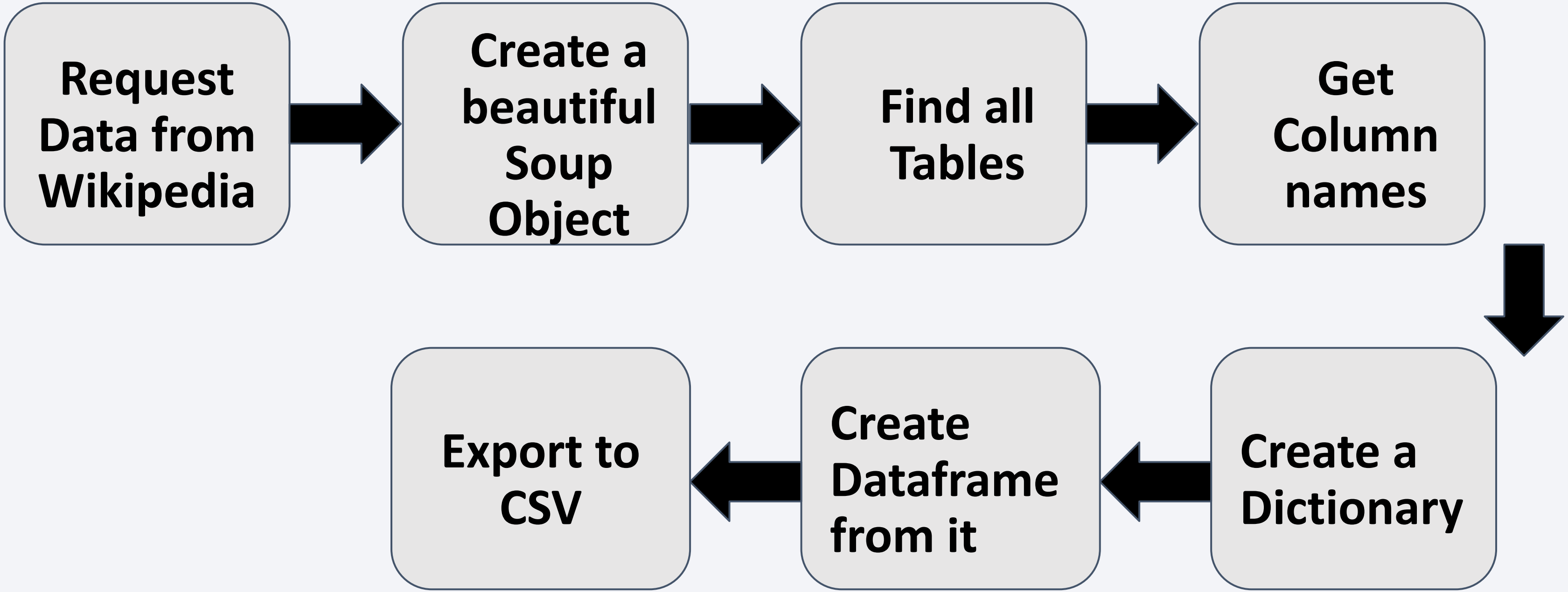
Data Collection – SpaceX API



Space X REST API URL: Link

GitHub URL: Link

Data Collection – Web Scraping



Wikipedia URL: Link

GitHub URL: Link

Data Wrangling

In the dataset, there are several cases where the booster did not land successfully.

- True Ocean, True RTLS, True ASDS means the mission has been successful.
- False Ocean, False RTLS, False ASDS means the mission was a failure.

We need to transform string variables into categorical variables where 1 means the mission has been successful and 0 means the mission was a failure.

1. Calculate launches number for each site

2. Calculate the number and occurence of each orbit

3. Calculate number and occurrence of mission outcome per orbit type

4. Create landing outcome label from Outcome column

5. Export to csv

EDA with Data Visualization

Charts Definitions:

- **Bar Charts:** Show the relationship between numeric and categoric values
- **Line Charts:** Line charts show trends in data over time
- **Scatter Charts:** Shows the relationship (correlation) between different variables

Bar Charts Plotted:
- Success rate vs Orbit

Line Charts Plotted:
- Success rate vs Year

Bar Charts Plotted:
- Flight Number vs Payload Mass
- Flight Number vs Launch Site
- Flight Number vs Orbit
- Payload Mass vs Orbit
- Payload vs Launch Site

Link: EDA (Data Viz)

EDA with SQL

SQL Performed Queries:

- Displaying the names of the unique lauunch 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 mas 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 records which will display the month names, faiilure landing_ouutcomes in drone ship, booster versions, launch_site for the
- months in year 2015.
- Rank the count of successful landiing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

Link: EDA SQL

Build an Interactive Map with Folium

The following items were created to improve the understanding within the data. They provide a clear overview of all launch sites, their surroundings, and the count of successful and unsuccessful landings, making the information easily accessible and comprehensible.

Markers of all Launch Sites:

- Red circle at NASA Johnson Space Center's coordinate with label showing its name
- Red circles at each launch site coordinates with label showing launch site name
- The grouping of points in a cluster to display multiple and different information for the same coordinates

Coloured Markers of the launch outcomes for each Launch Site:

- Markers to show successful and unsuccessful landings. Green for successful landing and Red for unsuccessful landing

Distances between a Launch Site to its proximities:

- Markers to show distance between launch site to key locations (railway, highway, coastway, city) and plot a line between them

Link: Interactive Map

Build a Dashboard with Plotly Dash

<u>Launch Site Dropdown:</u>

- Dropdown list to allows a user to choose the launch site or all launch sites

<u>Pie Chart:</u>

- Shows the total success and the total failure for the launch site chosen in the dropdown

<u>Rangeslider:</u>

- allows a user to select a payload mass in a fixed range

<u>Scatter Chart:</u>

- shows the relationship between two variables, in particular Success vs Payload Mass

<u>Plotly Dash URL:</u> Link

Predictive Analysis (Classification)

Data Preparation:

- Dataset loading
- Data standardization
- Data splitting into training and testing sets

Model Preparation:

- Selection of machine learning algorithms
- Setting parameters for each algorithm using GridSearchCV
- Training models using the GridSearchCV with the training dataset

Model Evaluation:

- Obtaining optimal hyperparameters for each model type
- Computing accuracy for each model using the test dataset
- Generating Confusion Matrix plots

Model Comparison:

- Evaluating models based on their accuracy
- Selection of the model with the highest accuracy

Predictive Analysis: Link

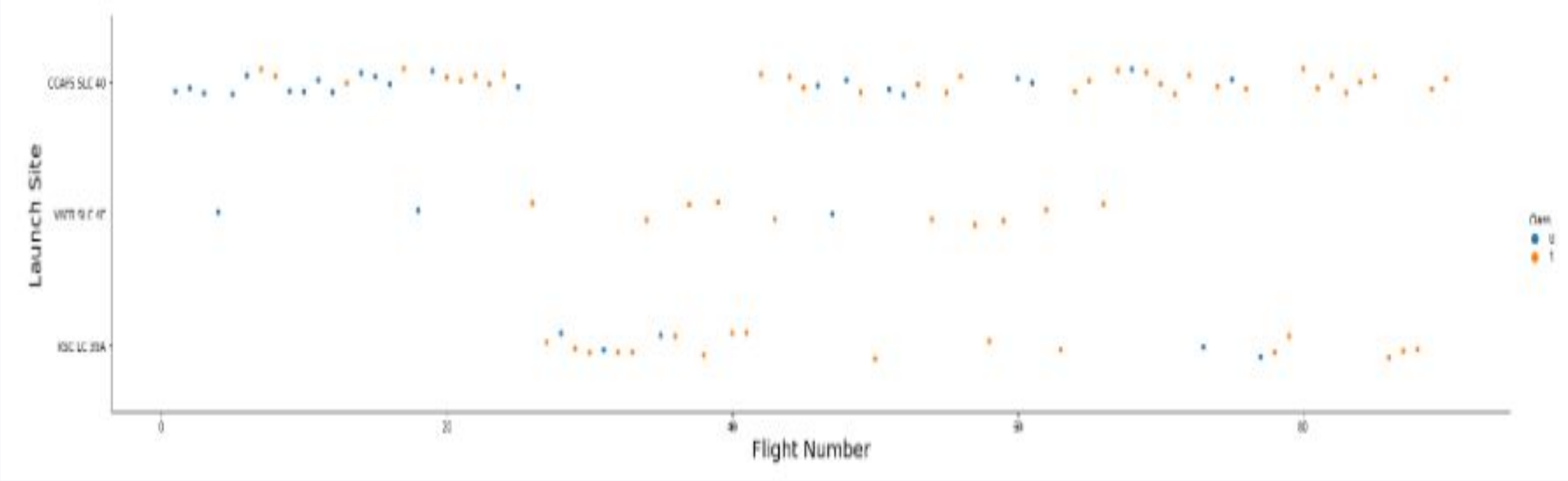
Results

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

Section 2

Insights drawn from EDA

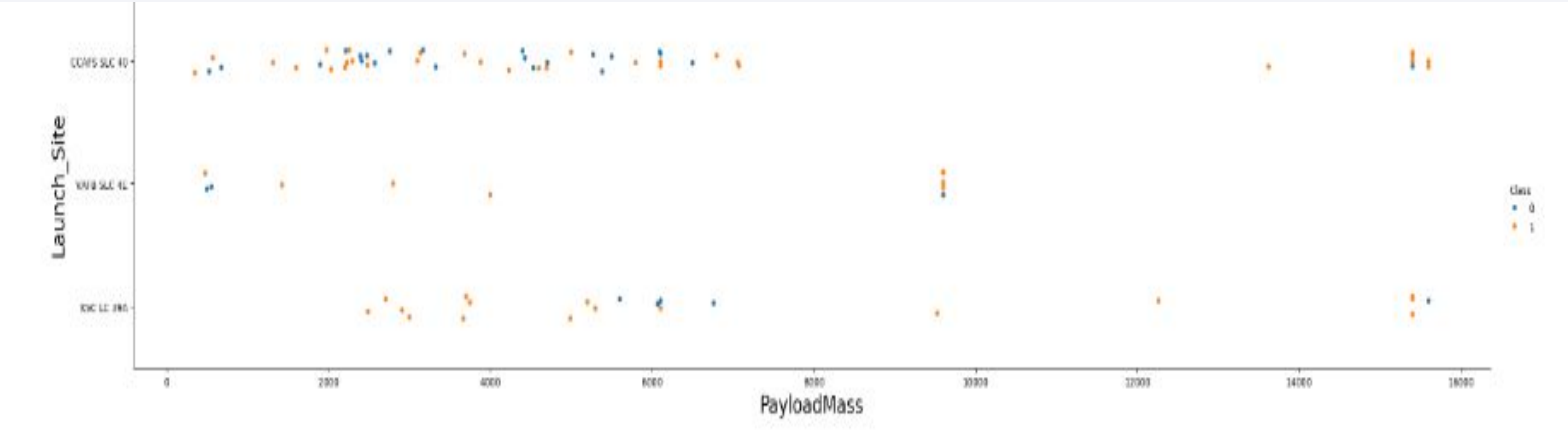
Flight Number vs. Launch Site



<u>Insights:</u>

- As the number of launches progressed, the success rate increased.
- VAFB SLC 4E and KSC LC 39A have higher success rates.

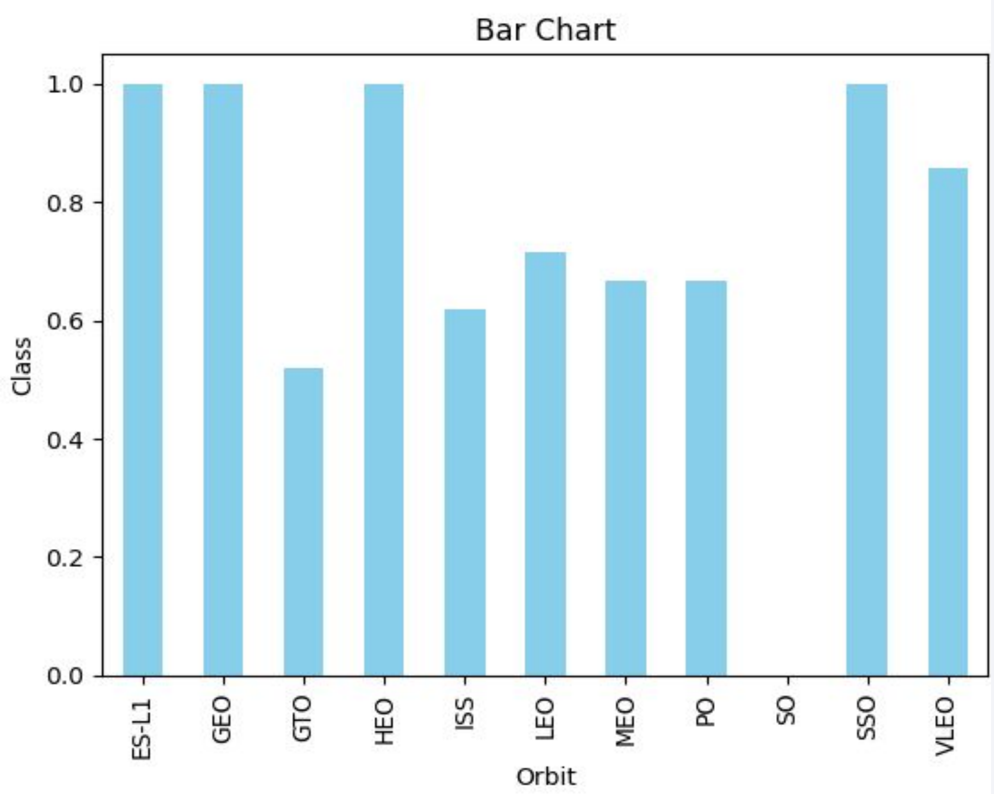
Payload vs. Launch Site



Insights:

- For each Launch Site, as the payload mass increases, the success rate also increases.
- KSC LC 39A maintains a 100% success rate even for payload masses under 5500 kg.

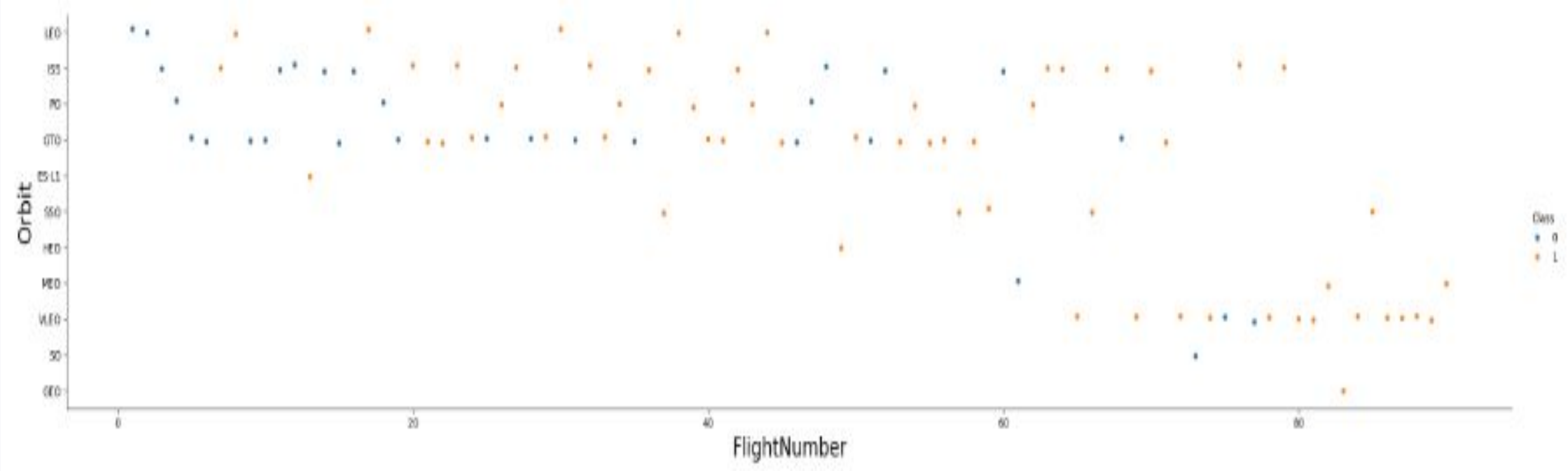
Success Rate vs. Orbit Type



Insights:

- **Orbital missions with a 100% success rate:** ES-L1, GEO, HEO, SSO
- .**Orbital missions with a 0% success rate:** SO
- **Orbital missions between 50% and 85% success rate:** GTO, ISS, LEO, MEO, PO

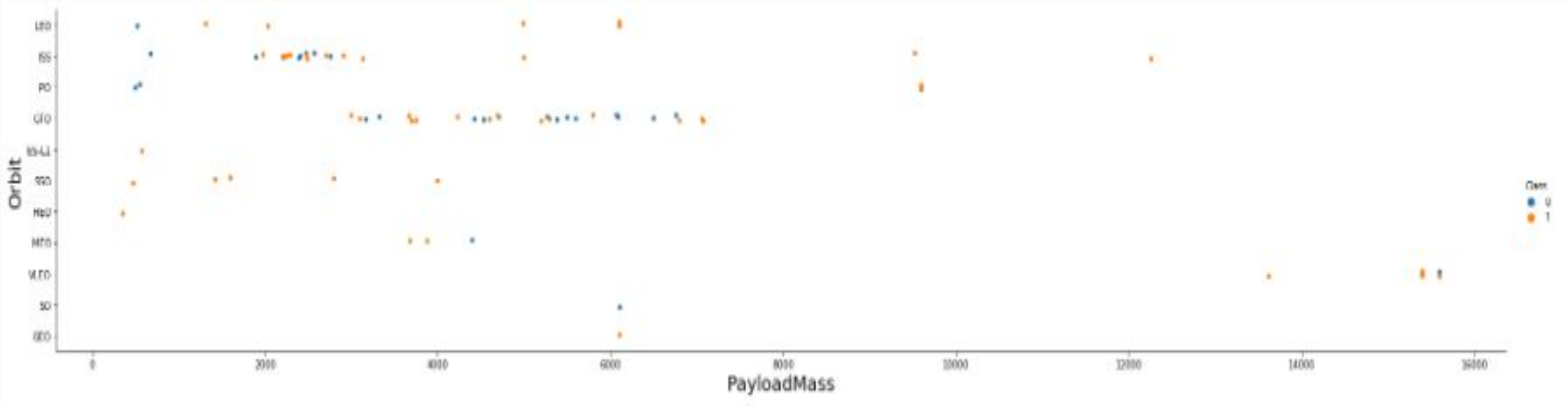
Flight Number vs. Orbit Type



Insights:

- For VLEO as the Flight Number increases, the success rate also increases.
- For some orbits like GTO, there is no relation between the success rate and the number of flights

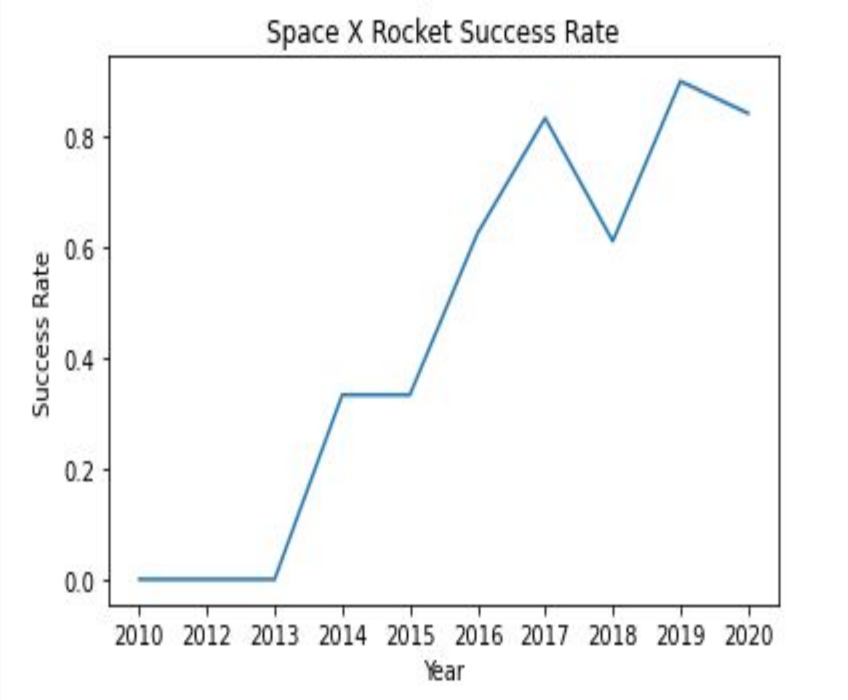
Payload vs. Orbit Type



Insights:

- Heavy payloads negatively affect Geostationary Transfer Orbits (GTO) but positively influence Geostationary Transfer Orbit (GTO) and Polar Low Earth Orbit (LEO) missions, such as those involving the International Space Station (ISS).

Launch Success Yearly Trend



Insights:

- For VLSince 2013, there has been an upward trend in the success rate of SpaceX rockets.

All Launch Site Names

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE
```

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

Insights:

The utilization of DISTINCT in the query helps eliminate duplicate LAUNCH_SITE entries.

Launch Site Names Begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site like "%CCA%" LIMIT 5
```

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	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	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

Insights:

- The WHERE clause, along with the LIKE clause, filters launch sites containing the substring CCA.
- The LIMIT 5 statement displays 5 records resulting from the filter.

Total Payload Mass

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS "NASA (CRS) Payload Mass"FROM SPACEXTABLE WHERE Customer = "NASA (CRS)"
```

NASA (CRS) Payload Mass
45596

Insights:

- The sum function returns the total payload mass.
- The WHERE clause filters the data to bring only the information corresponding to NASA (CRS).

Average Payload Mass by F9 v1.1

```
%sql select avg(PAYLOAD_MASS__KG_) AS "F9 v1.1 Avg Payload Mass" FROM SPACEXTABLE where Booster_Version = "F9 v1.1"
```

F9 v1.1 Avg Payload Mass
2928.4

Insights:

- The AVG() function returns the average payload mass as a result.
- The WHERE clause allows you to filter by the booster version.

First Successful Ground Landing Date

```
%sql SELECT min(Date) FROM SPACEXTABLE where Landing_Outcome = "Success (ground pad)"
```

min(Date)
2015-12-22

Insights:

- The MIN() function returns the earliest available date as a result.
- The WHERE clause allows you to filter by the first successful case.

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql SELECT Booster_Version FROM SPACEXTABLE where Landing_Outcome = "Success (drone ship)" and PAYLOAD_MASS__KG_ between 40
```

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

This query returns the booster version where landing was successful and payload mass is between 4000 and 6000 kg. The WHERE and AND clauses filter the dataset.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, COUNT(*) AS total_count FROM SPACEXTABLE GROUP BY mission_outcome
```

Mission_Outcome	total_count
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Listing the total number of successful and failure mission outcomes.

Boosters Carried Maximum Payload

```
[10]: %sql SELECT booster_version FROM spacextable WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM spacextable
```

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

Listing the names of the booster which have carried the maximum payload mass

2015 Launch Records

```
%sql SELECT substr(Date, 6, 2) AS month, Landing_Outcome, booster_version, launch_site FROM spacextable WHERE substr(Date, (
```

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

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

Listing the failed landing outcomes in drone ship, their booster versions and launch site names for the months in year 2015.

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

```
%sql SELECT Landing_Outcome, COUNT(*) AS count_of_outcomes FROM spacextable WHERE Date BETWEEN '2010-06-04' AND '2017-03-20
```

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

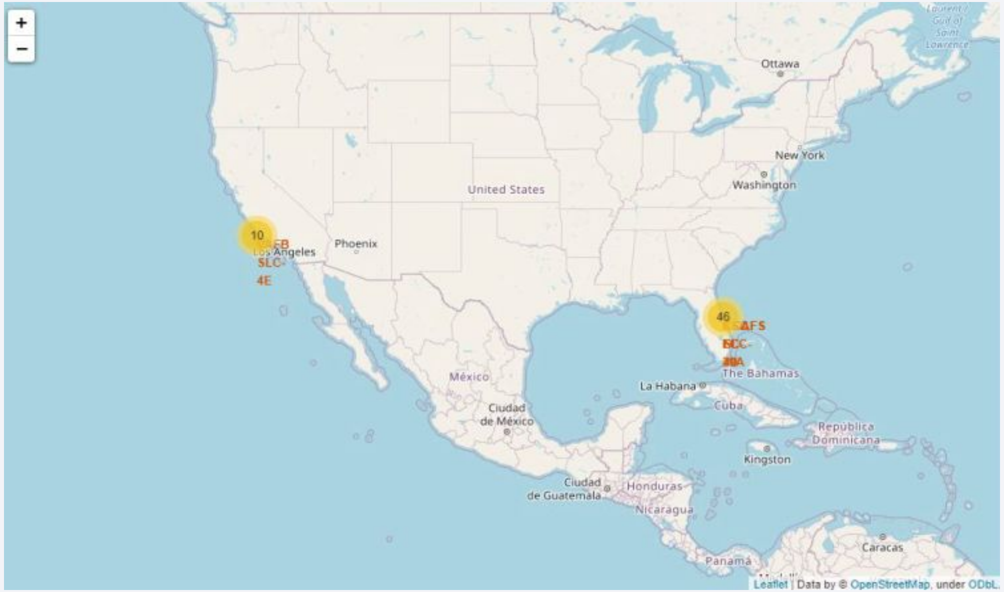
Landing_Outcome	count_of_outcomes
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Ranking the count of landing outcomes (such as Failure (droneship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20 in descending order.

Section 3

Launch Sites Proximities Analysis

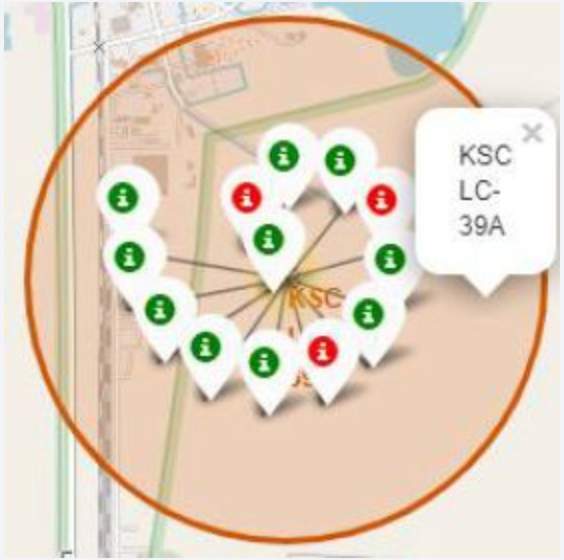
All Launch Sites Location Markers



Insights:

- All launch sites are situated near the coastline to mitigate the risk of debris falling or explosions occurring close to populated areas when rockets are launched towards the ocean.

Color Labeled Markers



Insights:

- All Green markers indicate successful launches.
- Red markers denote failed launches.
- Launch Site KSC LC-39A exhibits a notably high success rate.

Distances between CCAFS SLC-40 and its proximities

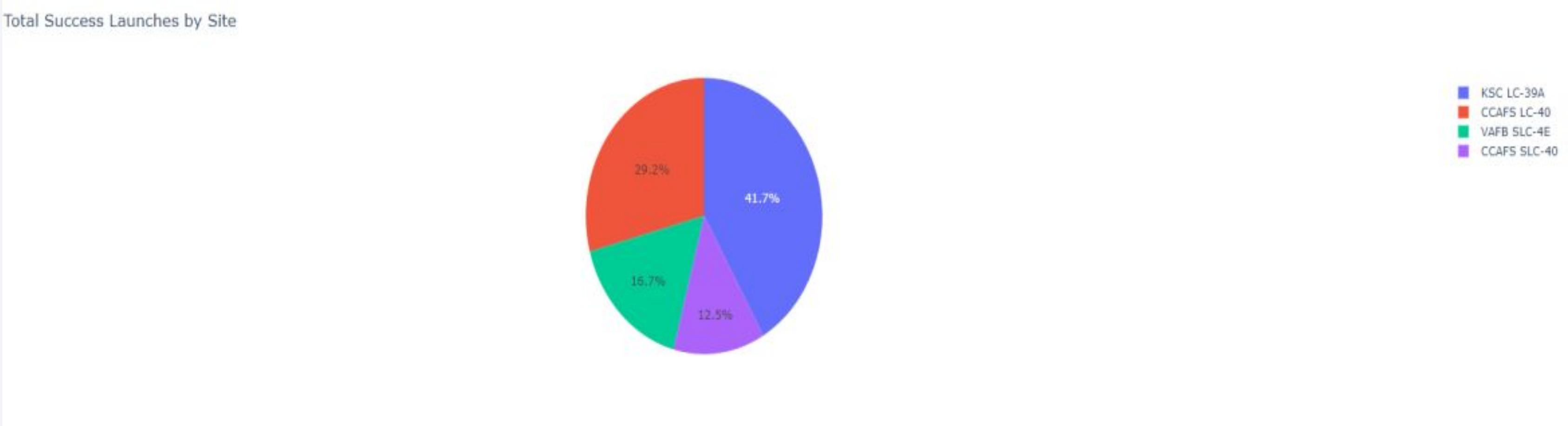
Insights:

The image shows that CCAFS SLC-40 is located 0.87 kilometers away from a coastline and 0.60km away from a railway.

Section 4

Build a Dashboard with Plotly Dash

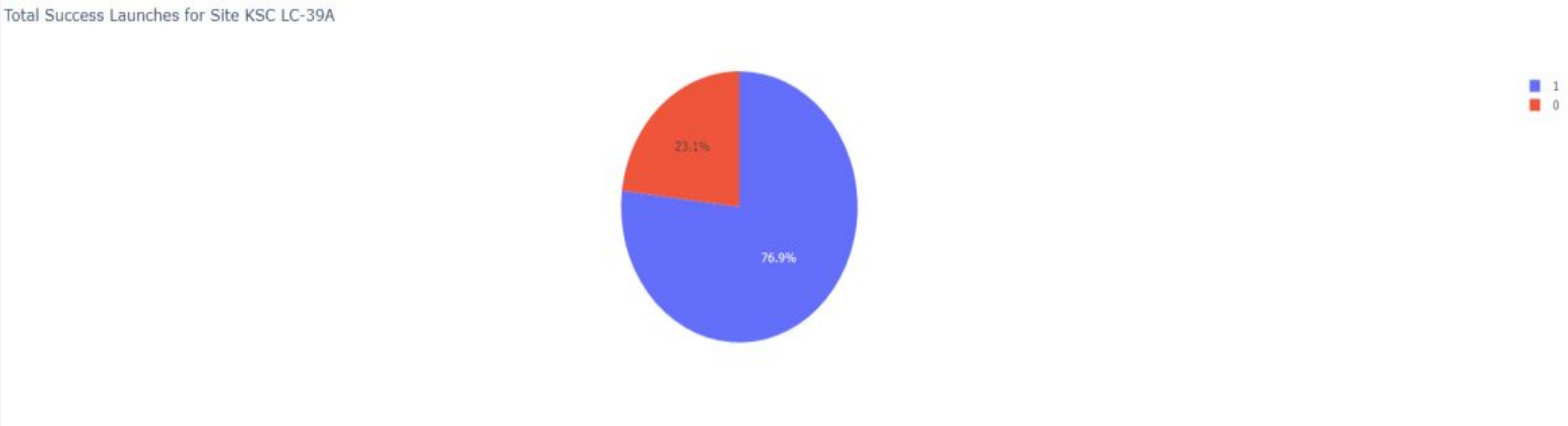
Total success by Site



Insights:

KSC LC-39A stands out in the chart as the launch site with the most successful launches among all.

Total success launches for Site KSC LC-39A



Insights:

With a launch success rate of 76.9%, KSC LC-39A boasts 10 successful landings and only 3 failures.

Payload mass vs Outcome for all sites

Insights:

The charts indicate that the highest success rate is observed for payloads between 2000 and 5500 kg.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

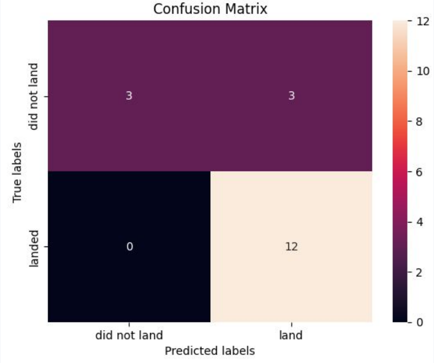
	Accuracy Train	Accuracy Test
Logreg	0.846429	0.833333
Svm	0.848214	0.833333
Tree	0.889286	0.833333
Knn	0.848214	0.833333

Insights:

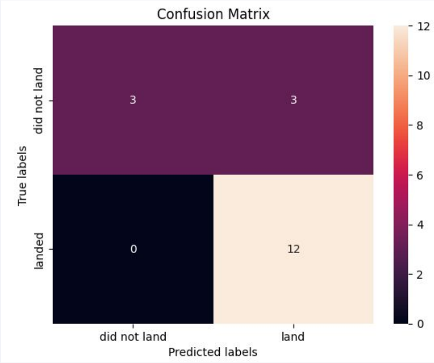
Based on the test set scores, the entire dataset affirm that the Decision Tree Model stands out as the best performer. Regarding accuracy, all methods performed similarly. To make a definitive choice, obtaining more test data would be beneficial.

Confusion Matrix

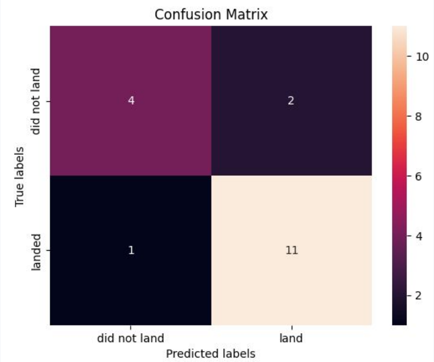
Log Reg



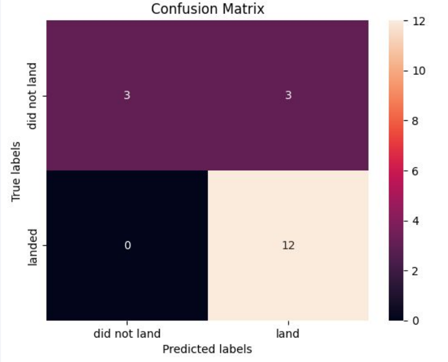
SVM



Decision Tree



KNN



Insights:

Since the test accuracies are identical, the confusion matrices also display uniformity across all models. The primary issue with these models lies in the occurrence of false positives.

Conclusions

The success of a mission can be attributed to various factors including the launch site, orbit type, and accumulated experience from previous launches. Notably, certain orbits such as GEO, HEO, SSO, and ES-L1 exhibit the highest success rates. Payload mass also plays a crucial role, with lighter payloads generally performing better across different orbits. Despite the dataset not providing clear insights into why certain launch sites, like KSC LC-39A, outperform others, obtaining additional atmospheric and relevant data could offer valuable explanations.

In conclusion, the Decision Tree Model emerges as the optimal algorithm for this dataset, primarily due to its superior train accuracy. Additionally, launches featuring lighter payload masses tend to yield better outcomes. Most launch sites are situated near the Equator line and coastline, potentially minimizing risks associated with debris and explosions. Over time, the success rate of launches has shown an upward trend, with KSC LC-39A boasting the highest success rate among all sites.

Thank you!