Applied Data Science Capstone

Kishan Bhakta
May 11, 2023

Outline

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

Executive Summary

Summary of Methods:

- Gathering data
- Organizing data
- Analyzing data through visualization
- Analyzing data using SQL
- Creating interactive maps with Folium
- Constructing dashboards with Plotly Dash
- Conducting predictive analysis (classification)

Summary of Findings:

- Findings from exploratory data analysis
- Screenshots demonstrating interactive analytics demo
- Results from predictive analysis

Introduction

Project Background and Context:

In the era of commercial space exploration, SpaceX has emerged as the leading company, revolutionizing the affordability of space travel. Their website showcases Falcon 9 rocket launches, priced at 62 million dollars, significantly lower than other providers whose costs exceed 165 million dollars per launch. The key factor behind SpaceX's cost advantage is the ability to reuse the first stage of their rockets. Therefore, by determining the likelihood of a successful first stage landing, we can estimate the overall cost of a launch. This project aims to leverage public information and machine learning models to predict whether SpaceX will reuse the first stage.

Questions to Address:

1. How do variables such as payload mass, launch site, number of flights, and orbits influence the probability of a successful first stage landing?

2. Is there an observable trend indicating an increase in the rate of successful landings over the years?

3. Which algorithm is best suited for binary classification in this particular case?

Methodology

Methodology for Data Collection:

- Utilizing the SpaceX Rest API for data retrieval
- Employing Web Scraping techniques on Wikipedia

Data Wrangling Steps Undertaken:

- Filtering and refining the collected data
- Addressing any missing values present
- Utilizing One Hot Encoding to prepare the data for binary classification

Exploratory Data Analysis (EDA) Conducted:

- Leveraging visualization techniques and SQL for in-depth analysis

Interactive Visual Analytics Performed:

- Utilizing Folium and Plotly Dash to create interactive visualizations

Predictive Analysis Carried Out:

- Developing, fine-tuning, and evaluating classification models to achieve optimal results

Data Collection

Data was collected using a combination of SpaceX's REST API and web scraping from Wikipedia. This allowed us to gather complete information on launches for a detailed analysis. The SpaceX REST API provided columns such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude. The web scraping from Wikipedia provided columns including Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, and Time.

Data Collection With The SpaceX API

1. Requesting rocket launch data from SpaceX API.

2. Decoding the response content using `.json()` and converting it into a dataframe using `.json_normalize()`.

3. Requesting the necessary launch information from SpaceX API by applying custom functions.

4. Organizing the obtained data into a dictionary.

5. Creating a dataframe from the dictionary.

6. Filtering the dataframe to include only Falcon 9 launches.

7. Replacing missing values in the Payload Mass column with the calculated mean for that column.

8. Exporting the data to a CSV file.

Data Collection Through Web Scraping

1. Requesting Falcon 9 launch data from Wikipedia.
2. Creating a BeautifulSoup object from the HTML response.
3. Extracting all column names from the HTML table header.
4. Collecting the data by parsing HTML tables.
5. Constructing the obtained data into a dictionary.
6. Creating a dataframe from the dictionary.
7. Exporting the data to a CSV file.

Data Wrangling

Within the dataset, there are various scenarios where the booster did not achieve a successful landing. These instances include both attempted landings that failed due to accidents and specific outcomes based on the landing location.

For example:

- "True Ocean" indicates a successful landing in a designated region of the ocean.
- "False Ocean" denotes an unsuccessful landing in a designated region of the ocean.
- "True RTLS" signifies a successful landing on a ground pad.
- "False RTLS" indicates an unsuccessful landing on a ground pad.
- "True ASDS" represents a successful landing on a drone ship.
- "False ASDS" indicates an unsuccessful landing on a drone ship.

To facilitate analysis and modeling, these outcomes are primarily converted into training labels, where a value of "1" signifies a successful booster landing, and a value of "0" indicates an unsuccessful landing.

Data Visualization EDA

Plotted charts include:

1. Flight Number vs. Payload Mass
2. Flight Number vs. Launch Site
3. Payload Mass vs. Launch Site
4. Orbit Type vs. Success Rate
5. Flight Number vs. Orbit Type
6. Payload Mass vs. Orbit Type
7. Success Rate Yearly Trend

Scatter plots were utilized to visualize the relationships between variables, which can be useful in machine learning models.

Bar charts were used to compare discrete categories, aiming to demonstrate the relationship between specific categories and a measured value.

Line charts were employed to showcase trends in data over time, particularly for time series analysis.

SQL EDA

- Retrieve the names of unique launch sites in the space mission.
- Display 5 records where launch sites start with the string 'CCA'.
- Calculate the total payload mass carried by boosters launched by NASA (CRS).
- Determine the average payload mass carried by booster version F9 v1.1.
- List the date when the first successful landing outcome on a ground pad was achieved.
- List the names of boosters that successfully landed on a drone ship with a payload mass greater than 4000 but less than 6000.
- List the total number of successful and failed mission outcomes.
- Identify the booster versions that carried the maximum payload mass.
- List the failed landing outcomes on a drone ship, including their booster versions and launch site names, for the months in the year 2015.
- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the dates 2010-06-04 and 2017-03-20 in descending order.

Folium Interactive Map

Markers for all Launch Sites:

- Created a Marker with a Circle, Popup Label, and Text Label for NASA Johnson Space Center, utilizing its latitude and longitude coordinates as the starting location.
- Added Markers with Circles, Popup Labels, and Text Labels for all Launch Sites, indicating their geographical positions and proximity to the Equator and coastlines.

Colored Markers for launch outcomes at each Launch Site:

- Incorporated colored Markers, such as Green for successful launches and Red for failed launches, using Marker Cluster to identify Launch Sites with relatively high success rates.

Distances between a Launch Site and its surroundings:

- Included colored Lines to illustrate the distances between a Launch Site (e.g., KSC LC-39A) and nearby features like Railways, Highways, Coastlines, and the Closest City.

Plotly Dashboard

Launch Sites Dropdown List:

- Implemented a dropdown list to facilitate the selection of Launch Sites.

Pie Chart depicting Success Launches (All Sites/Certain Site):

- Created a pie chart to visually represent the total count of successful launches for all sites, and if a specific Launch Site was chosen, to display the breakdown of success and failure counts for that site.

Slider for Payload Mass Range:

- Integrated a slider component to enable the selection of a desired Payload mass range.

Scatter Chart illustrating Payload Mass vs. Success Rate for different Booster Versions:

- Utilized a scatter chart to showcase the relationship between Payload mass and Launch Success, specifically focusing on different Booster Versions.

Classification Model

1. Creating a NumPy array from the "Class" column in the data.

2. Standardizing the data using StandardScaler, including fitting and transforming the data.

3. Splitting the data into training and testing sets using the train_test_split function.

4. Creating a GridSearchCV object with a cross-validation value of 10 to identify the best parameters.

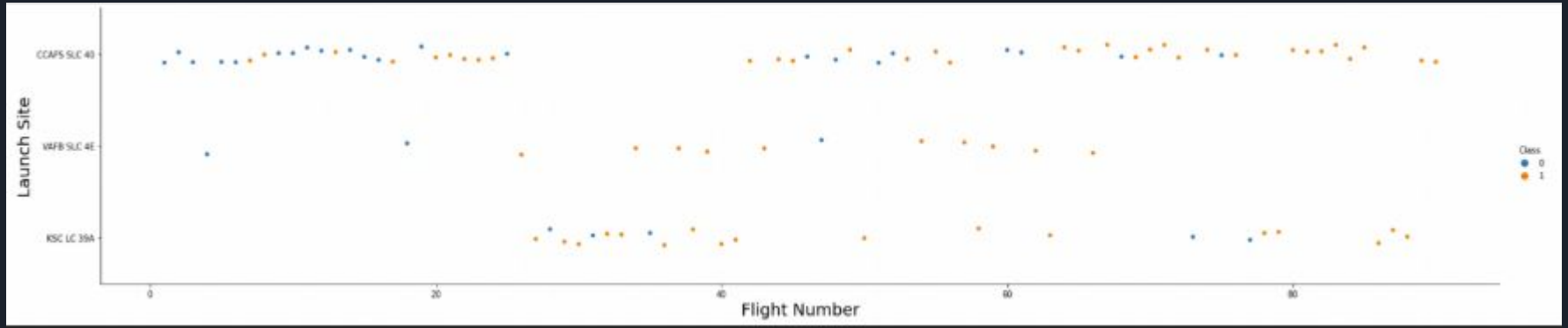
5. Applying GridSearchCV on Logistic Regression (LogReg), Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN) models.

6. Calculating the accuracy on the test data using the .score() method for all models.

7. Examining the confusion matrix for each model.

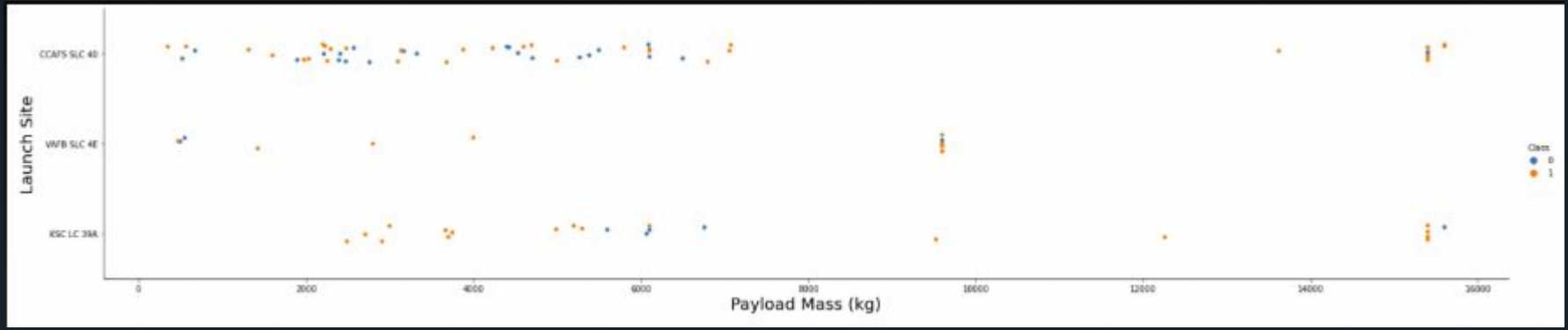
8. Determining the best-performing method by assessing the Jaccard_score and F1_score metrics.

Flight Number Vs. Launch Site



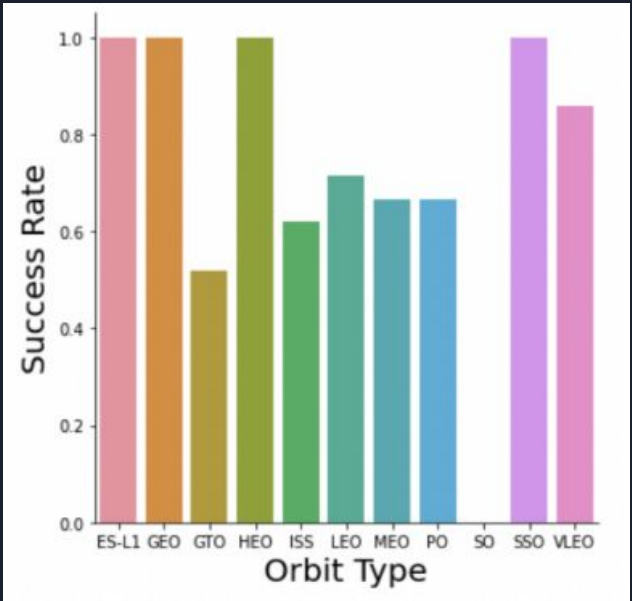
- The initial flights resulted in failures, whereas the most recent flights achieved success consistently.
- Approximately half of all launches took place at the CCAFS SLC 40 launch site.
- VAFB SLC 4E and KSC LC 39A exhibit higher success rates compared to other launch sites.
- An assumption can be made that the success rate of each subsequent launch tends to improve.

Payload Vs. Launch Site



- Across all launch sites, there is a positive correlation between payload mass and success rate, meaning that higher payload masses tend to result in higher success rates.
- The majority of launches with a payload mass exceeding 7000 kg achieved success.
- Additionally, KSC LC 39A boasts a 100% success rate for payload masses under 5500 kg as well.

Success Rate Vs. Orbit Type



Orbits with a 100% success rate:

- ES-L1
- GEO
- HEO
- SSO

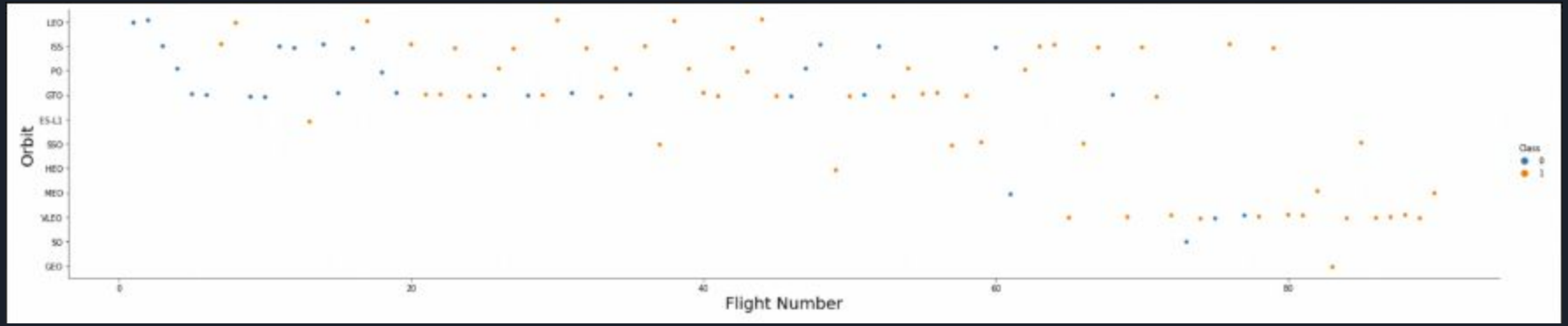
Orbits with a 0% success rate:

- SO

Orbits with a success rate between 50% and 85%:

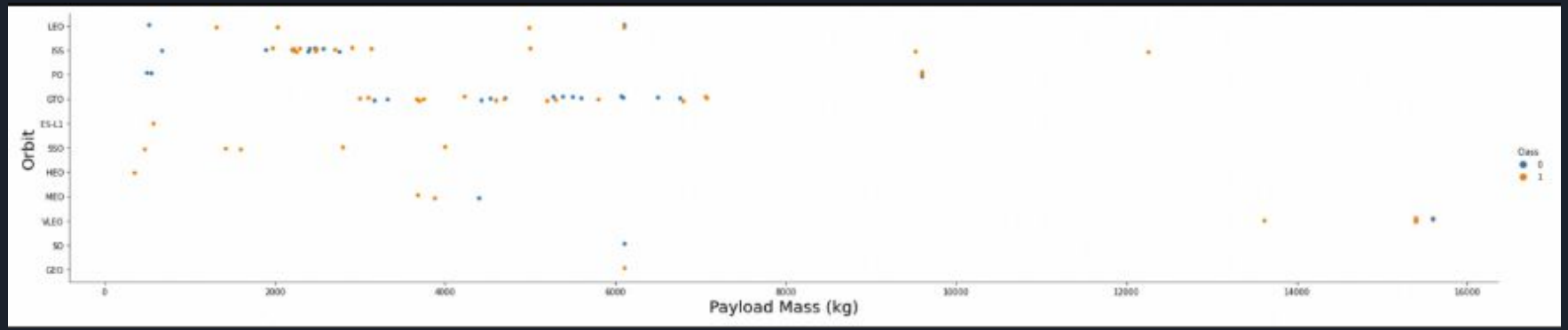
- GTO
- ISS
- LEO
- MEO
- PO

Flight Number Vs. Orbit



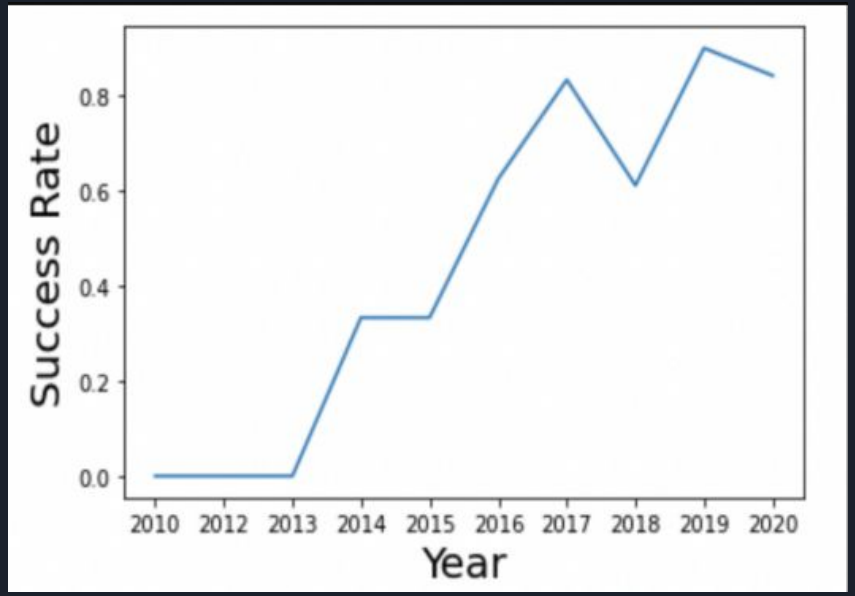
In the LEO orbit, the success rate appears to be associated with the number of flights, suggesting a relationship. However, in the GTO orbit, there seems to be no discernible correlation between flight number and success.

Payload Mass Vs. Orbit Type



Heavy payloads exert a negative impact on GTO orbits but have a positive influence on GTO and Polar LEO (ISS) orbits.

Launch Success Trends



The success rate has exhibited a continuous increase from 2013 to 2020.

Launch Site Names

```
In [4]: %sql select distinct launch_site from SPACEXDATASET;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[4]:

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

Showing the unique launch site names involved in the space mission.

CCA Launch Sites

```
In [5]: %sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[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	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

Showing 5 records where the launch sites start with the string 'CCA'.

Sum Payload Mass

```
In [6]: %sql select sum(payload_mass__kg_) as total_payload_mass from SPACEXDATASET where customer = 'NASA (CRS)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[6]:

total_payload_mass
45596

Showing the cumulative payload mass carried by boosters launched by NASA (CRS).

F9 v1.1 Average Payload Mass

```
In [7]: %sql select avg(payload_mass__kg_) as average_payload_mass from SPACEXDATASET where booster_version like '%F9 v1.1%';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[7]:

average_payload_mass
2534

Showing the average payload mass carried by the booster version F9 v1.1.

First Date For Successful Ground Landing

```
In [8]: %sql select min(date) as first_successful_landing from SPACEXDATASET where landing__outcome = 'Success (ground pad)';
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[8]:

first_successful_landing
2015-12-22

Listing the date of the initial successful landing outcome on a ground pad.

Drone Ship Landing With Payload Between 4000 & 6000

```
In [9]: %sql select booster_version from SPACEXDATASET where landing__outcome = 'Success (drone ship)' and payload_mass__kg_ between 4
        000 and 6000;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[9]:

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

Listing the names of the boosters that have achieved success on a drone ship and have a payload mass greater than 4000 but less than 6000.

Successful & Failed Mission Outcomes

```
In [10]: %sql select mission_outcome, count(*) as total_number from SPACEXDATASET group by mission_outcome;
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[10]:

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

Providing a list of the total number of successful and failed mission outcomes.

Boosters Carrying The Maximum Payload

```
In [11]: %sql select booster_version from SPACEXDATASET where payload_mass__kg_ = (select max(payload_mass__kg_) from SPACEXDATASET);
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[11]:

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 versions that have carried the highest payload masses.

2015 Launch Records

```
In [12]: %%sql select monthname(date) as month, date, booster_version, launch_site, landing__outcome from SPACEXDATASET
         where landing__outcome = 'Failure (drone ship)' and year(date)=2015;

 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90108kqblod8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[12]:

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Listing the drone ship landing outcomes that resulted in failure, along with their corresponding booster versions and launch site names, for the months in the year 2015.

Rank Success Count From 2010/06/04 - 2017/03/20

```
In [13]: %%sql select landing__outcome, count(*) as count_outcomes from SPACEXDATASET
         where date between '2010-06-04' and '2017-03-20'
         group by landing__outcome
         order by count_outcomes desc;
```

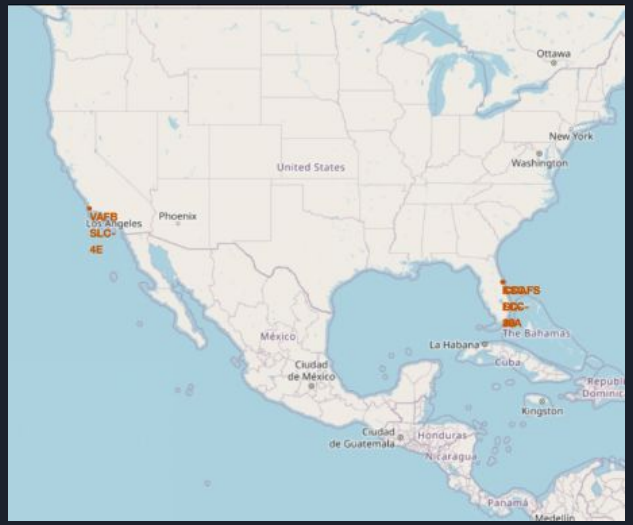
```
 * ibm_db_sa://wzf08322:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[13]:

landing__outcome	count_outcomes
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

Ranking the number of landing outcomes (e.g., Failure on drone ship or Success on ground pad) in descending order between the dates 2010-06-04 and 2017-03-20.

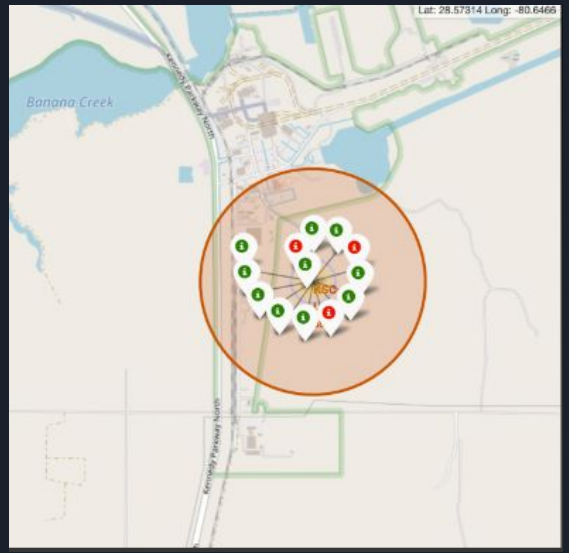
Launch Site Location Markers On Map



The majority of Launch sites are located near the Equator, where the Earth's surface moves faster compared to other regions. When a spacecraft is launched from the Equator, it maintains this initial speed and continues to orbit around the Earth. This phenomenon is a result of inertia and assists the spacecraft in maintaining a sufficient speed for successful orbit.

Additionally, all launch sites are strategically situated in close proximity to the coast. Launching rockets towards the ocean reduces the risk of debris falling or exploding near populated areas, ensuring safety.

Color Labeled Launch Records On Map

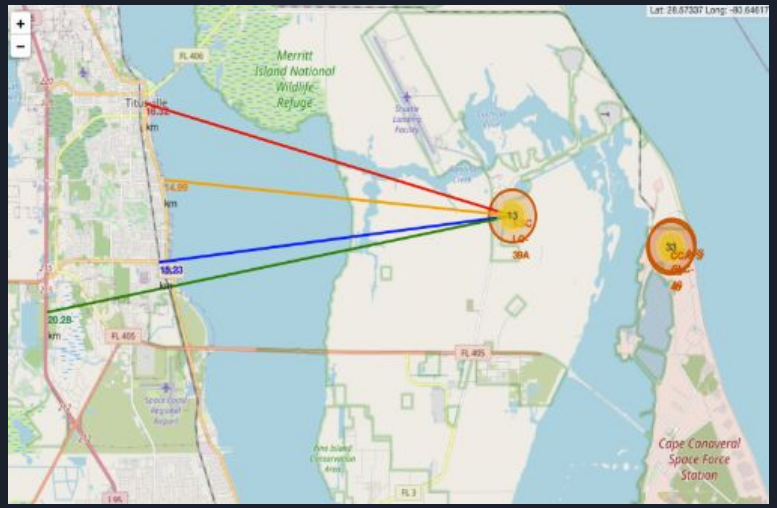


By observing the color-coded markers, it should be straightforward to identify launch sites with relatively high success rates.

- Green markers indicate successful launches.
- Red markers indicate failed launches.

Furthermore, Launch Site KSC LC-39A demonstrates an exceptionally high success rate.

Distance From KSC LC -39A To Its Proximities

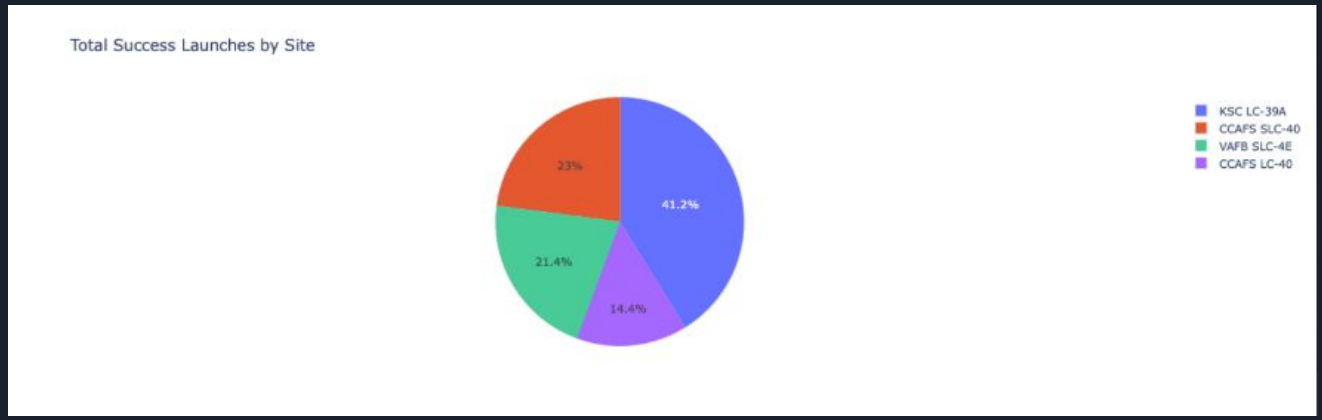


Through visual analysis of Launch Site KSC LC-39A, the following observations can be made:

- The site is in relative proximity to a railway, approximately 15.23 km away.
- It is also relatively close to a highway, located about 20.28 km away.
- Additionally, the site is near the coastline, at a distance of approximately 14.99 km.
- Furthermore, Launch Site KSC LC-39A is in close proximity to its nearest city, Titusville, at a distance of around 16.32 km.

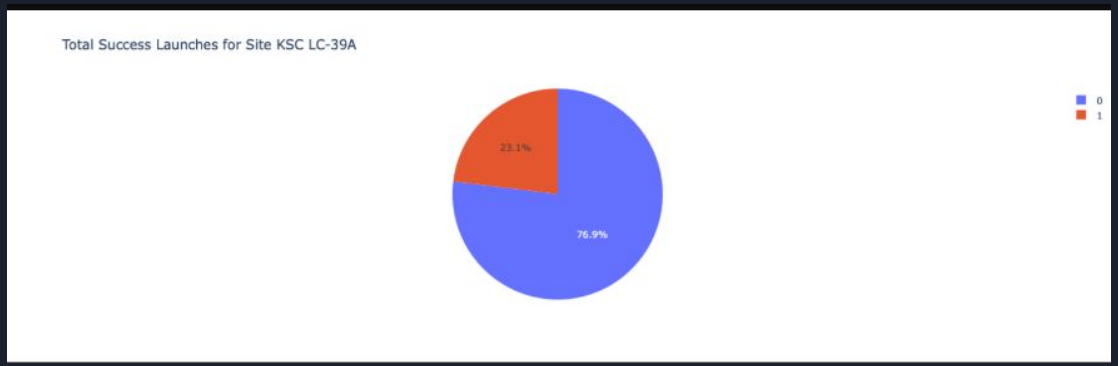
Considering the high speed of failed rockets, covering distances of 15-20 km within seconds, it poses potential risks to populated areas.

Launch Success Count For All Sites



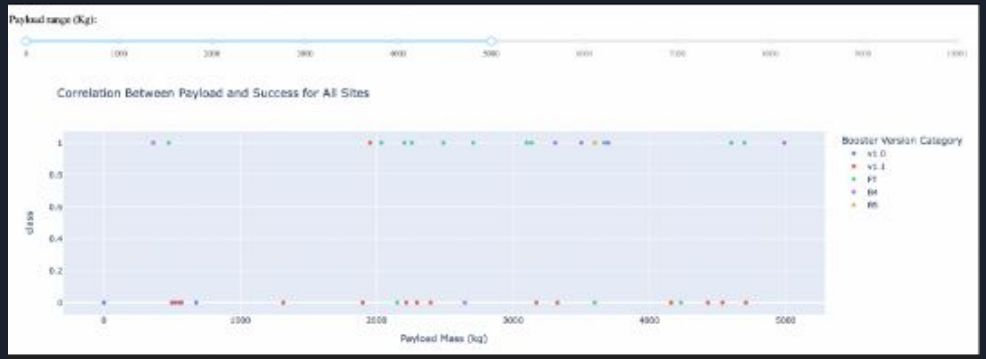
The chart provides clear evidence that among all the launch sites, KSC LC-39A stands out with the highest number of successful launches.

Launch Site With The Highest Success Ratio



KSC LC-39A boasts the highest launch success rate, standing at 76.9%, with a total of 10 successful landings and only 3 failed landings.

Payload Mass Vs. Launch Outcomes



The charts indicate that payloads ranging from 2000 kg to 5500 kg exhibit the highest success rate.

Classification Accuracy

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

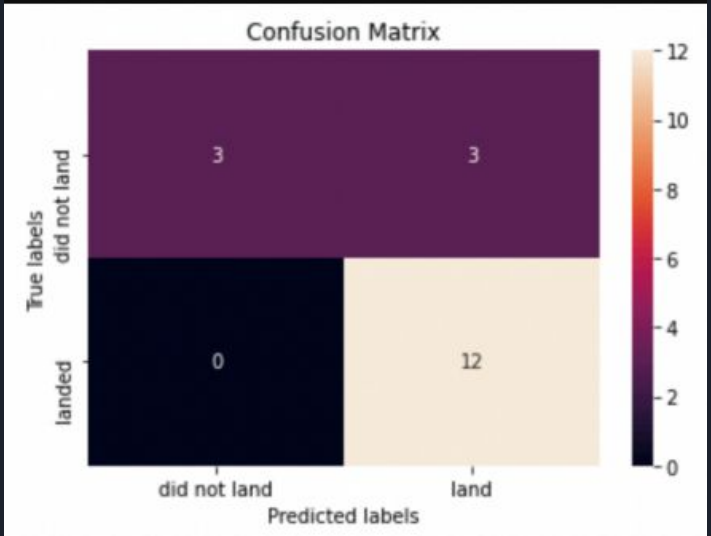
	LogReg	SVM	Tree	KNN
Jaccard_Score	0.833333	0.845070	0.882353	0.819444
F1_Score	0.909091	0.916031	0.937500	0.900763
Accuracy	0.866667	0.877778	0.911111	0.855556

Upon evaluating the scores of the Test Set (Top Data), it is inconclusive to determine the best-performing method.

The similarity in Test Set scores could be attributed to the small sample size (18 samples). To obtain a more accurate assessment, we conducted tests on the entire Dataset.

Considering the scores of the complete Dataset (Bottom Data), it is evident that the Decision Tree Model emerges as the best model. This model not only yields higher scores but also exhibits the highest accuracy among the methods evaluated.

Confusion Matrix



By analyzing the confusion matrix, it becomes evident that logistic regression effectively differentiates between the various classes. However, the main issue lies in the occurrence of false positives, which poses a significant challenge.

Conclusion

- The Decision Tree Model proves to be the most effective algorithm for this dataset.
- Launches with lower payload masses exhibit more favorable outcomes compared to those with larger payload masses.
- The majority of launch sites are situated near the Equator line, and all sites are in close proximity to the coast.
- The success rate of launches demonstrates an upward trend over the years.
- Among all the launch sites, KSC LC-39A exhibits the highest success rate.
- Orbits ES-L1, GEO, HEO, and SSO demonstrate a 100% success rate.

Appendix

- GitHub Link
  - https://github.com/SirKB1/DataScienceCapstone
- Thank You To All Of The Instructors:
  - IBM Data Science Professional Certificate | Coursera
- Wikipedia
- SpaceX Dataset For Rocket Launches