Space Y

First Stage, Falcon 9 reusage analysis

Juraj Žigo

April 30, 2024

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
  - Visualization – Charts
  - Dashboard
- Discussion
  - Findings & Implications
- Conclusion

EXECUTIVE SUMMARY

My approach:

- Data Collection
- Data Wrangling
- Exploratory Data Analysis
- Data Visualization
- Building a Dashboard with Plotly Dash
- Predictive Analysis
- Conclusion

INTRODUCTION

- The company’s Space Y goal is to rival the price of the competition by reusing its first stage and avoiding incremental costs to each flight,
- The aim of the project is to predict if the Falcon 9 first stage will land successfully and hence can justify the price which is some 60% lower than the competition,
- Reusing the first stage is dependent on safe landing,
- Historical data are available to be able to evaluate and find optimal parameters for the future flights. Available data to be analyzed are as follows:
  - Date & Time
  - Launch site
  - Payload Mass
  - Orbit
  - Mission Outcome
- The final report will conclude on the prerequisites for the future for the first stage to land safely and hence, justify the price to the customers.

METHODOLOGY

- Data Collection – API
  - Requesting rocket launch data from SpaceX API using URL -> json file parsing conversion to pandas dataframe -> creating a subset with selected features -> restricting the dataset to the booster version Falcon 9 only
- Data Collection – Web scrapping
  - Requesting booster version Falcon9 from Wikipedia -> Creating a Beautiful Soup object from HTML -> extracting all column names from the HTML doc <Header> -> parsing the HTML table -> creating a dictionary -> creating a pandas dataframe from the dictionary -> data exporting to the csv file
- Data Wrangling
  - Checking the missing value using isnull().sum() method -> replacing missing values with the mean using .mean() method -> exporting the data_falcon9.to_csv method -> Exploratory Data Analysis -> determination of labels for further trainings -> calculation of launches per each site -> calculate the number and occurrence per each orbit -> calculate the number of mission outcomes per each orbit -> creating the landing outcome label -> exporting the data into a csv file
- Exploratory Data Analysis
  - **Data visualization** – scatter plots : Flight Number vs Payload, Flight No vs Launch Site, Payload Mass vs launch Site, Orbit Type vs Success Rate, Flight No vs Orbit Type, Payload Mass vs Orbit Type, Success Rate Yearly trend
  - **Scatter plots -** help to visually identify causalities, a linear or other type of relationships between variables. Features with some kind of correlation can be further used for machine learning processing.
  - **Bar plots** – are used to compare discrete categories and identifying the significance of items being compared.
  - **Line charts -** show trends in data over the x-axis, time series

METHODOLOGY

- Exploratory Data Analysis
  - **SQL –** displaying the names of unique launch sites | displaying 5 records where launch sites name start with 'CCA' | displaying total payload mass carried by boosters launched by NASA | displaying avg payload mass carried by the booster Falcon9 v1.1 | listing dates when the first successful landing outcome in ground pad was achieved | listing the names of boosters which have success in drone ship and have payload mass between 4-6 tons | |listing the total number of successful and failure mission outcomes |listing the names of the booster versions which have carried the max payload mass | listing failure outcomes in drone ship, booster version and launch site names during months in 2015| ranking the count of landing outcomes – Failure - drone ship or Success – ground pad from June 4, 2010 – to March 20, 2017
- Building an Interactive map with Folium
  - Markers of all Launch Sites: added marker with circle | popup label and text label of NASA Johnson space centre using its latitude and longitude coordinates as a start location | added markers with circle, popup label and text label of all launch sites - latitude and longitude coordinates to show proximity to equator and coasts. coloured markers of the launch outcomes | success : green and failed :red | launches using marker cluster to identify which launch sites have relatively high success rates | distances between a launch site to its proximities | coloured lines to show distances between the launch site KSC LC-39A | further proximities like railway, highway, coastline and closest city
- Building a Dashboard with Dash Plotly
  - Launch Sites Dropdown List | added a dropdown list to enable Launch Site selection | created a pie chart showing Success Launches | added a pie chart to show the total successful launches count for all sites and the success vs. failed counts for the site | added a slider to select payload range | scatter chart of payload mass vs. success rate for the different booster versions: | scatter plot to show the correlation between payload and launch success
- Predictive analysis – process and classifiers:
  - Creating a Numpy array from 'Class' data -> Normalizing data with StandardScaler -> Train_Test_Split -> Creating a GridSearchCV object with cv=10 to find the best parameters -> Applying GridSearchCV on all models listed below -> Calculating accuracy on the test data: *score()* method -> Creating a confusion matrix for all models -> Evaluating the best score for the final model selection. ML models evaluated were as follows:
    - KNN
    - Logistic Regression
    - Decision Tree
    - Support Vector Machine
- Results and conclusion are shown in the following slides

DATA COLLECTION

- Data Set obtained using the API
- Conversion of data from JSON to Pandas data frame
- Restricting the data related to the booster version Falcon 9 only, the results are as follows:

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Bl
0	1	2006-03-24	Falcon 9	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	N
1	2	2007-03-21	Falcon 9	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	N

DATASET EXAMINATION

After having processed the raw data from the previous step, I looked into the data set to get the initial insights before we move to the predictive analysis.

SQL queries were used to retrieve the data from the database to further analyze relationships between the data as follows:

Quantitative:

- Launch sites
- Number of launches per site
- payload mass carried by boosters
- payload mass carried by booster version F9 v1.1
- Successful landings
- Number of successful and failure missions
- Booster version data

EDA with SQL

Display names of unique launches in the space mission

Display the names of the unique launch sites in the space mission

```
[9]: #%sql SELECT * FROM SPACEXTABLE LIMIT 3;
     %sql SELECT DISTINCT Launch_Site FROM SPACEXTABLE ;

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

[9]:

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

Display 5 records where launch sites begin with the string 'CCA'

Display 5 records where launch sites begin with the string 'CCA'

```
[23]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE '%CCA%' LIMIT 2;

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

[23]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parach
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 (parach

EDA with SQL

Display the total payload mass carried by boosters launched by NASA (CRS)

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
[34]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS 'Total Payload' FROM SPACEXTABLE;
```

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

[34]:

Total Payload
619967

Average payload mass carried by booster version F9 v1.1.

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql SELECT ROUND(AVG(PAYLOAD_MASS__KG_),3) AS 'AVG_MASS_BOOSTER F9_v1.1' FROM SPACEXTABLE;
```

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

AVG_MASS_BOOSTER F9_v1.1
6138.287

EDA with SQL

List the names of the boosters which have success in drone ship and have payload mass 4-6 tons

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql SELECT MIN(Date) AS 'First Successfull Landing Date' FROM SPACEXTABLE WHERE Mission_Outcome='Success';
```

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

First Successfull Landing Date
2010-06-04

List the names of boosters which have success in drone ship and have payload mass 4-6 tons

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

```
%sql SELECT DISTINCT Booster_Version FROM SPACEXTABLE WHERE Mission_Outcome = 'Success' AND Landing_Outcome LIKE '%dro
```

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

Booster_Version
F9 FT B1020
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

IBM Developer

SKILLS NETWORK

EDA with SQL

List the total number of successful and failure mission outcomes

Task 7 ¶

List the total number of successful and failure mission outcomes

```
%sql SELECT DISTINCT Mission_Outcome, COUNT(*) AS Total FROM SPACEXTABLE GROUP BY Mission_Outcome;
```

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

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

List the names of boosters which have carried the maximum payload mass.

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql SELECT Booster_Version,PAYLOAD_MASS__KG_ FROM SPACEXTABLE  WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

EDA with SQL

List the records which will display the month names, failure landing outcomes in drone ship, booster versions, launch site for the months in 2015

Task 9

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 6,2) as month to get the months and substr(Date,0,5)='2015' for year.

```
%sql SELECT substr(Date, 6, 2) AS Month, Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTABLE WHERE substr(D
```

```
 * 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

Rank the count of landing outcomes (such as Failure(drone ship) or Success (ground pad) between 20, in descending order)

Task 10

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

```
0]: %sql SELECT Landing_Outcome, COUNT(*) AS Outcome_Count FROM SPACEXTABLE WHERE Date BETWEEN '2010-06-04' AND '2017-03
```

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

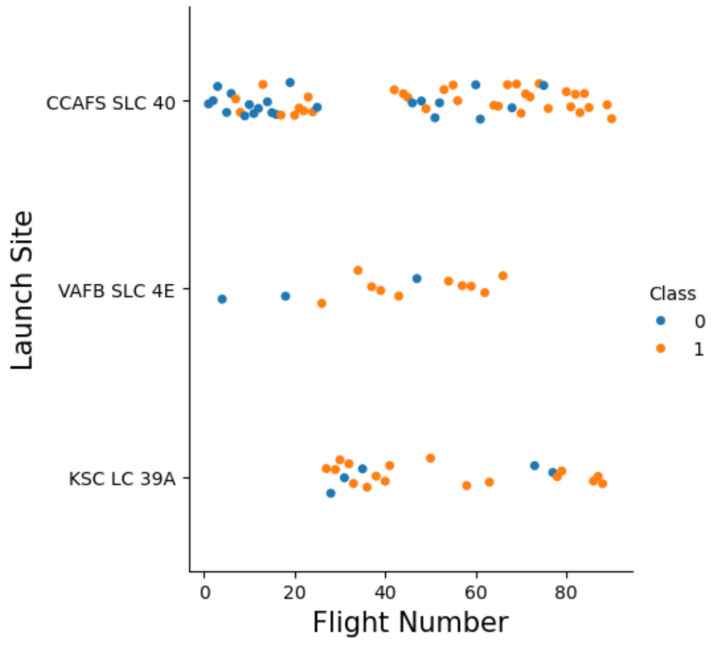
0]:

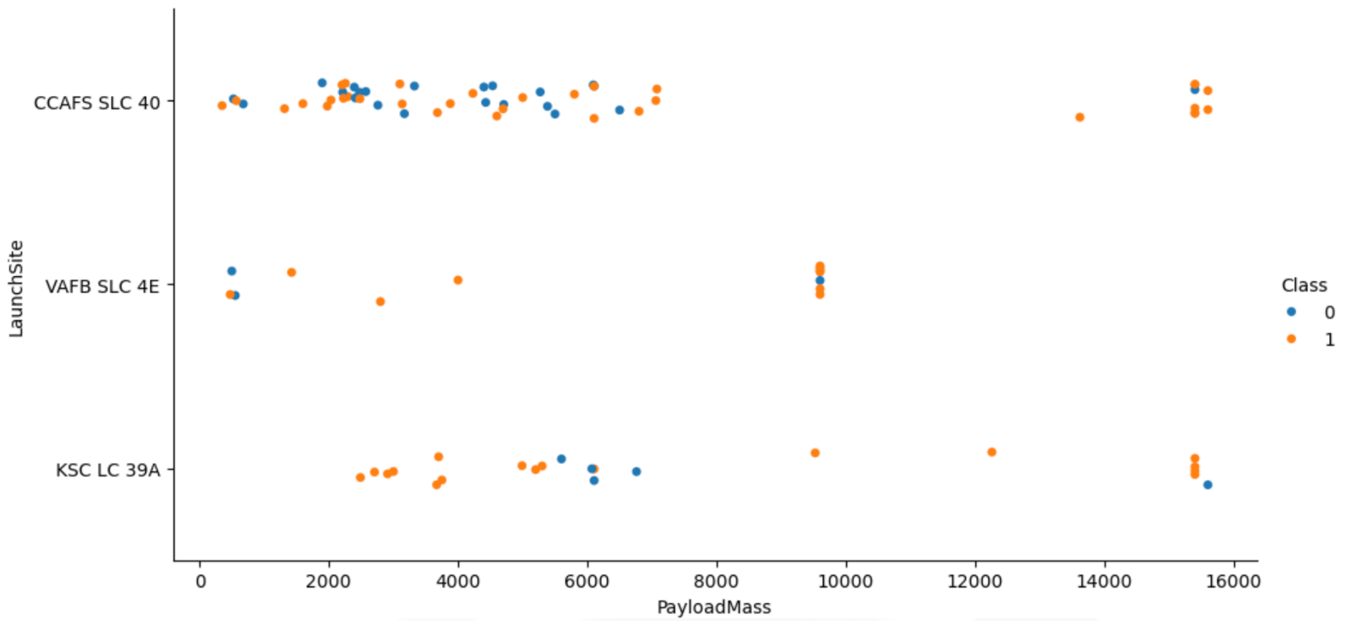
Landing_Outcome	Outcome_Count
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

VISUAL EXAMINATION OF THE DATA

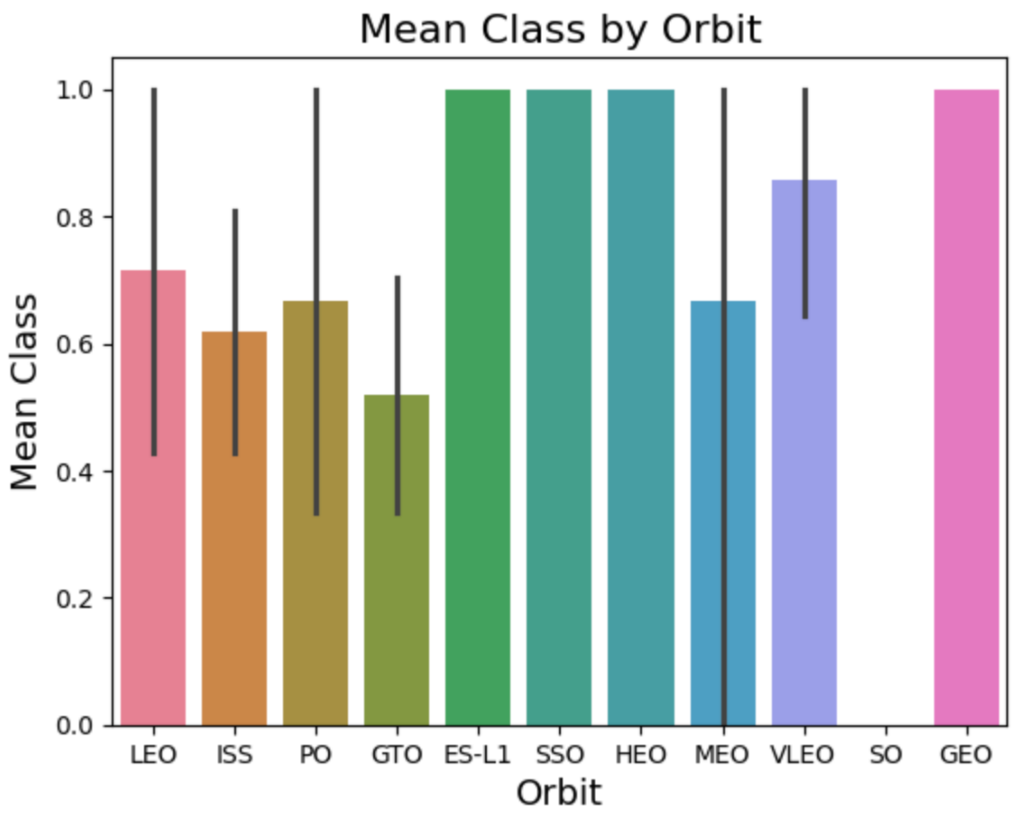
THE NEXT STEP IS TO IDENTIFY CAUSALITIES IN DATA BY USING VARIOUS CHARTS. I will also share my insights in each data visualization.

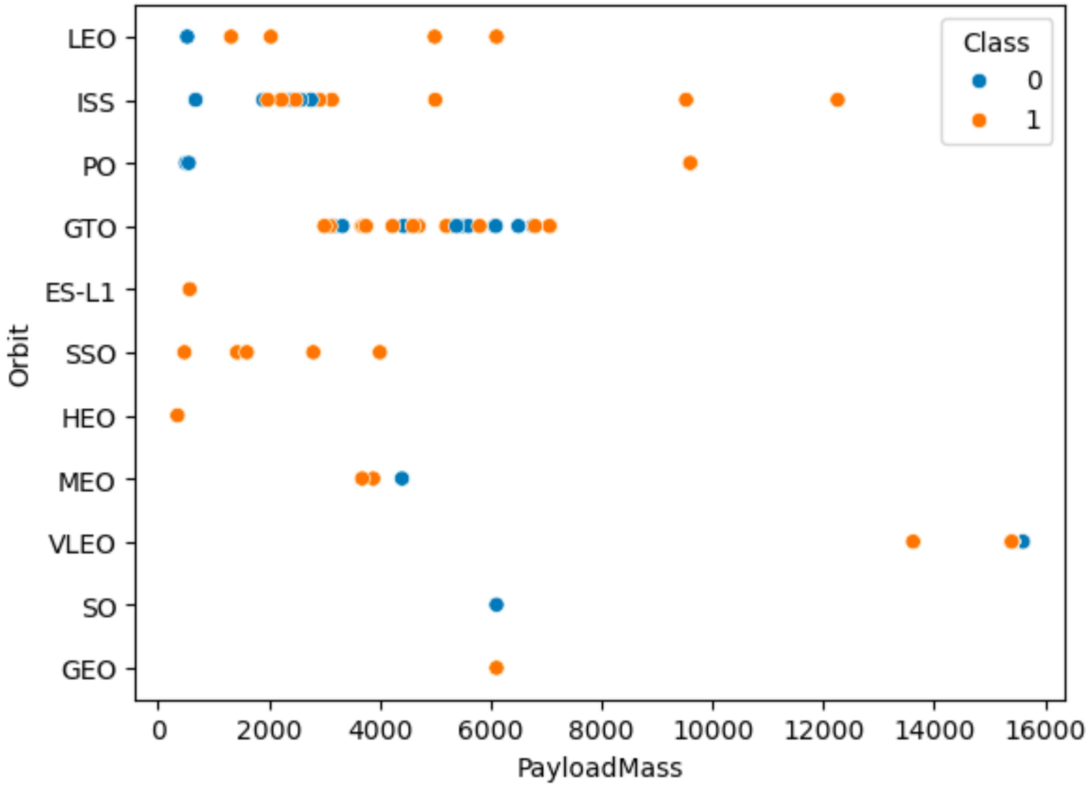
While the launch site SLC40 has higher number of launches, LC39A is more successful



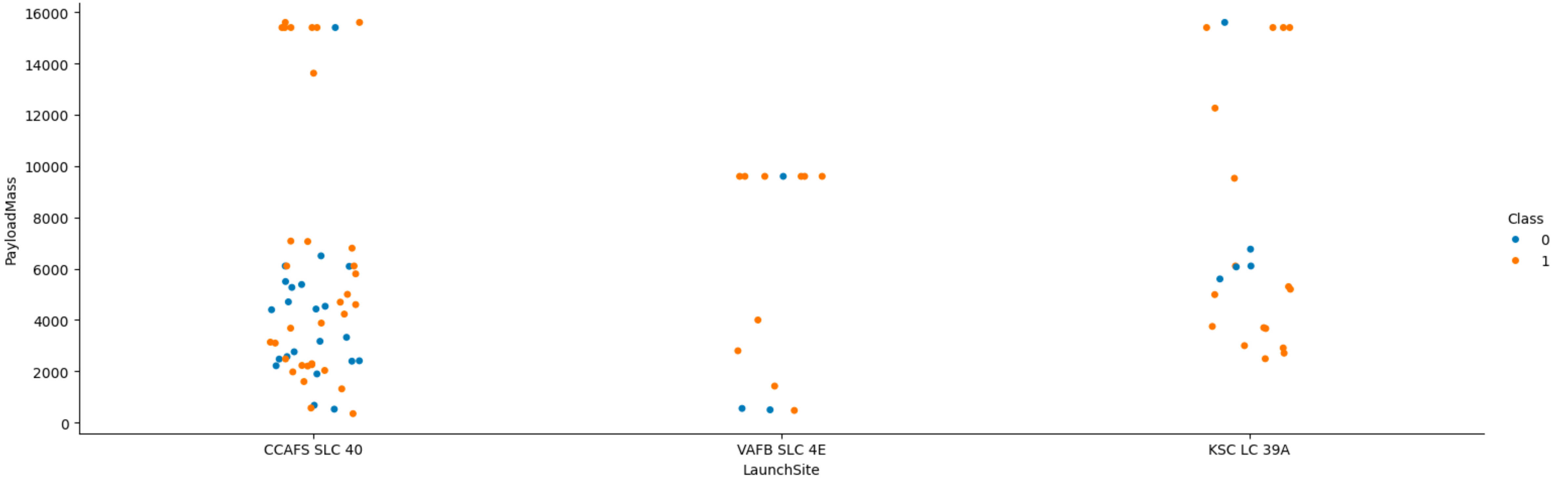


The Orbit 'SSO' is the safe bet due to 100% success rate and highest number of launches

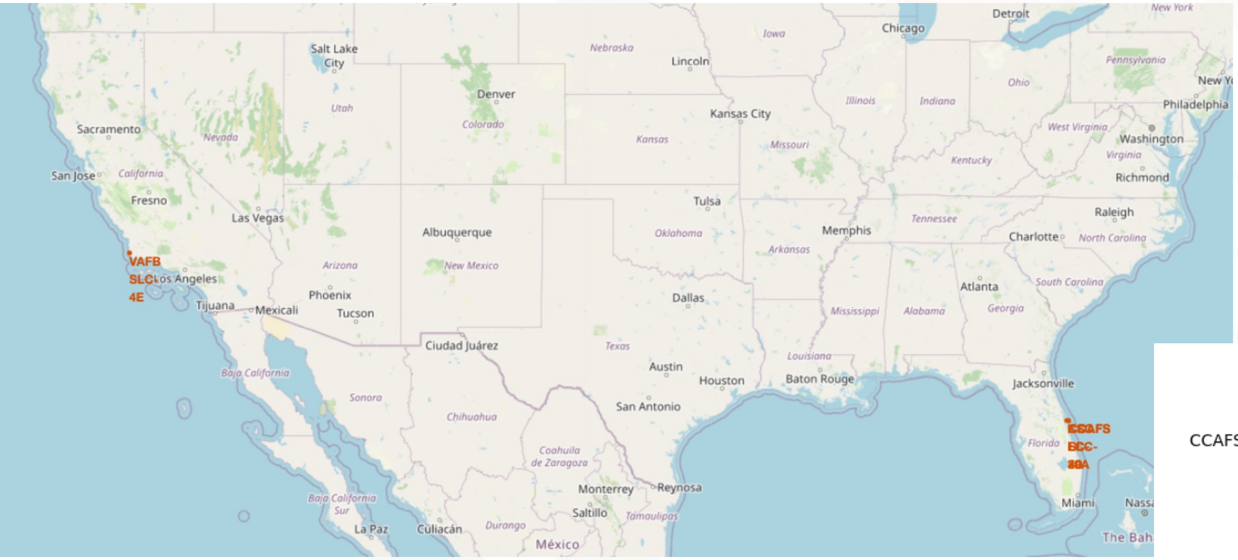


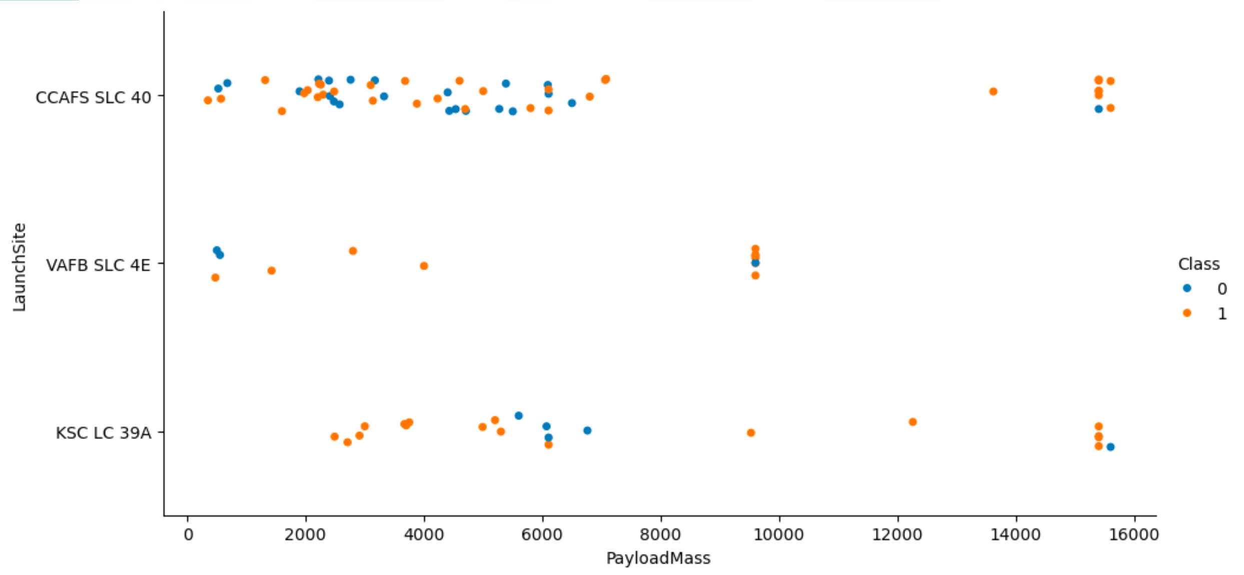


When the PayLoad Mass exceeds 10 tons, success rate increases dramatically. Lighter PayLoad Mass below 4 tons works well in LC39A which however not consistent in all launch sites

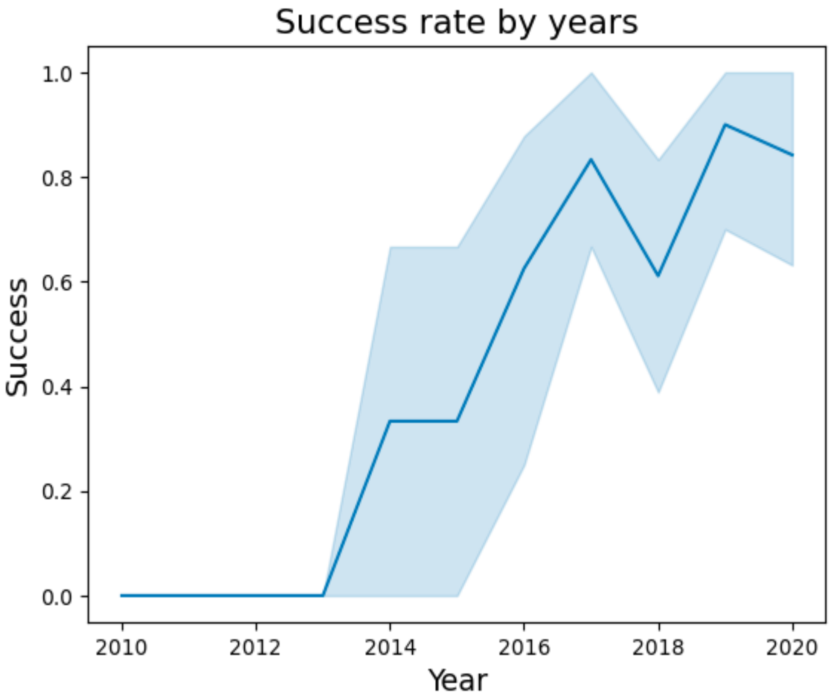


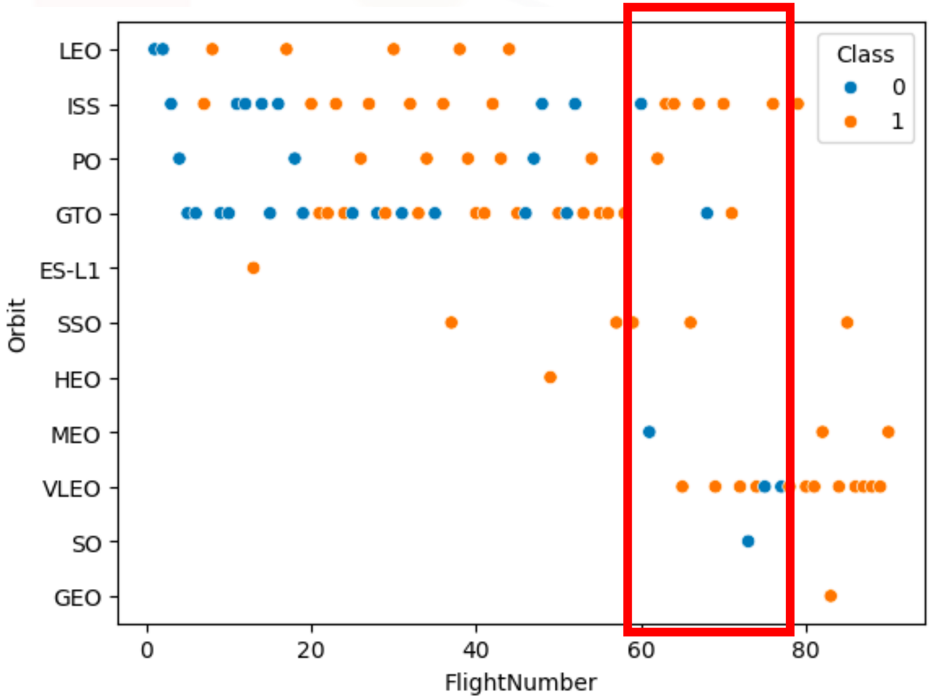
All 3 Launch sites have proximity to the Equator and coastline. However, those in Florida, show a higher number of launches.



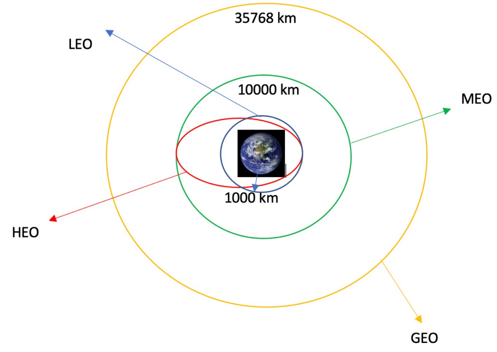


Success rate improves with time with the exception of the year 2018 which show the experimentation with various orbits



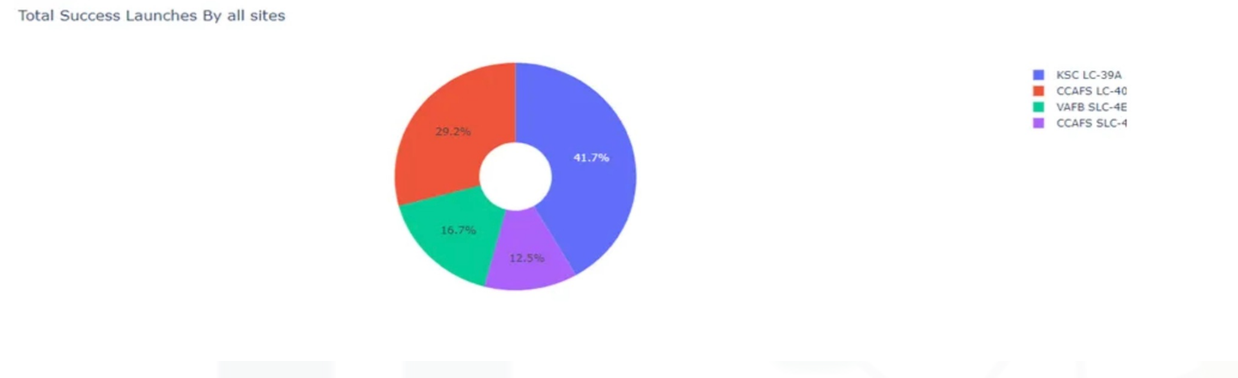


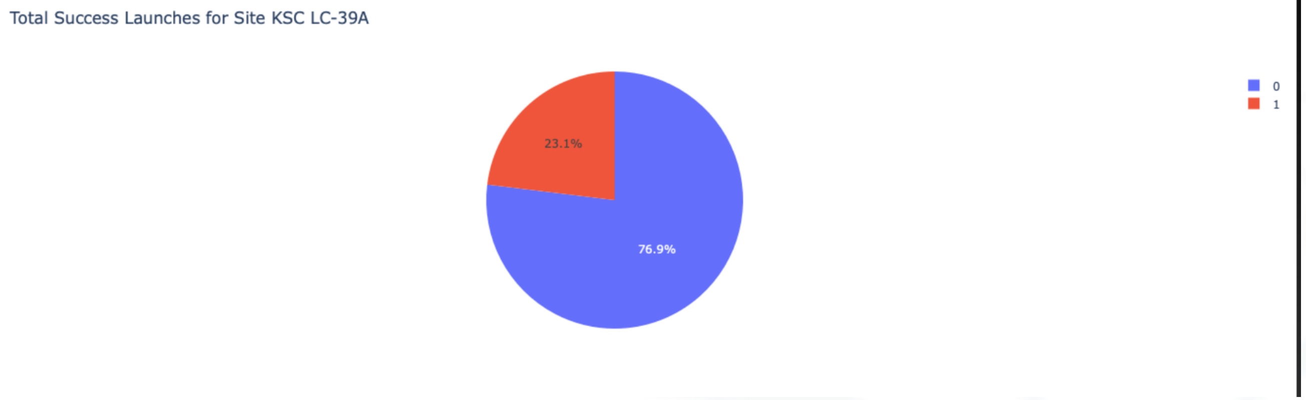
Highest number of orbits in timeframe



Orbits

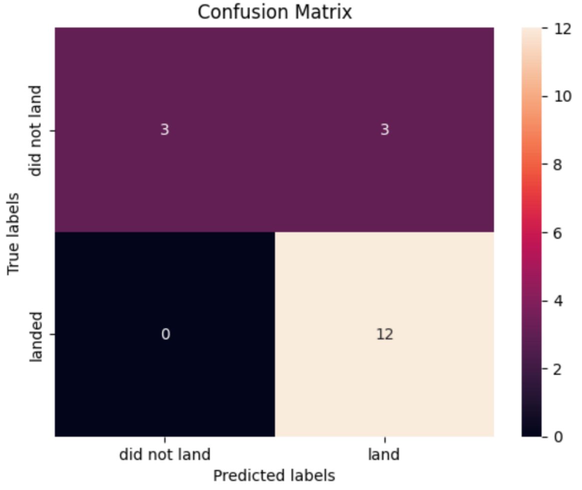
The site LC39A shows a 41% share on total successful launches with its success rate being 77%





The test accuracy of the models are similar across the board, resulting in 83% with the exception of Decision Trees which show only 55% accuracy.

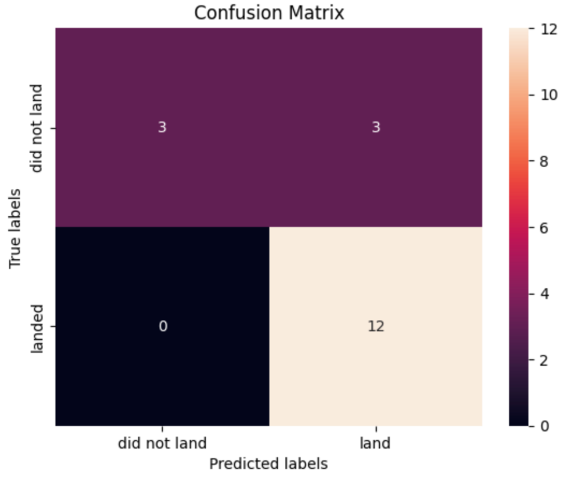
Logarithmic Regression




```
# Print the test accuracy
print("Test Accuracy:", test_accuracy_log)
```

```
Test Accuracy: 0.8333333333333334
```

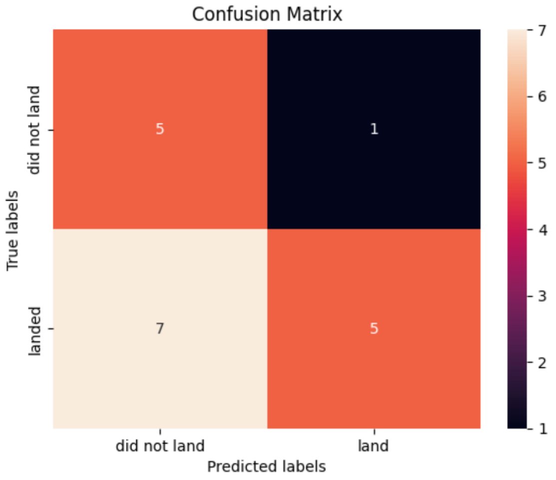
Support Vector Machine




```
# Print the test accuracy
print("Test Accuracy:", test_accuracy_svm)
```

```
Test Accuracy: 0.8333333333333334
```

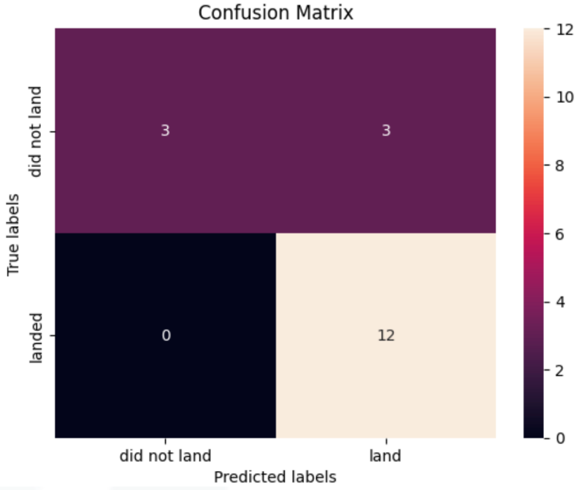
Decision Trees




```
# Print the test accuracy
print("Test Accuracy:", test_accuracy_tree)
```

```
Test Accuracy: 0.5555555555555556
```

KNN




```
# Print the test accuracy
print("Test Accuracy:", test_accuracy_knn)
```

```
Test Accuracy: 0.8333333333333334
```

We can plot the confusion matrix

CONCLUSION

In order to maximize success rate of the first stage reusability, the following shall be adhered to:

- **Launch site** should be **KSC LC 39A** – highest share of 41% with its success rate of 77%. It's close to the equator which is also ideal and cost efficient for launching to the space.
- Use the **orbit SSO** given 100% success rate and the highest number of launches. Other orbits (LEO, ES-L1 and HEO) showing 100% have a small number of launches and might be riskier than SSO
- With the **PayLoad Mass** exceeds 10 tons, success rate increases dramatically. Lighter PayLoad Mass below 4 tons works well only at the site KSC LC 39A
- All machine learning models show equally 83% accuracy with the exception of Decision Trees which show only 55%. More data need to be collected and applied machine learning models re-run again and tested of their accuracy to be used in the future.