IBM Data Science Capstone Project

SPACEX

RANA SARI

12/9/2023

Outline

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

Executive Summary

- Summary of methodologies

1. Data Collection
2. Data Wrangling
3. EDA with Data Visualization
4. EDA with SQL
5. Building an Ineractive Map with Folium
6. Building a Dashboard with Plotly Dash
7. Predictive Analysis (Classification)

- Summary of all results

1. Exploratory Data Analysis Results
2. Interactive Analytics Demo in Screenshots
3. Predictive Analysis Results

Introduction

- Project background and context

We predicted whether the Falcon 9's first stage would land successfully. SpaceX promotes Falcon 9 rocket launches on its website at a cost of 62 million dollars, a significant savings compared to other providers who charge upwards of 165 million dollars. Much of this cost reduction is attributed to SpaceX's ability to reuse the first stage. Therefore, determining the success of the first stage landing allows us to estimate the launch cost. This information is valuable for potential competitors bidding against SpaceX for rocket launches.

- Problems you want to find answers

1.What factors influence the successful landing of the rocket?
2.What is the impact of each relationship with specific rocket variables on determining the success rate of a landing?
3.What conditions does SpaceX need to achieve to attain the best results and ensure the highest success rate for rocket landings?

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Rest API
- Perform data wrangling
  - One Hot Encoding data fields for Machine Learning and dropping irrelevant columns
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Select and train a suitable algorithm, tune its hyperparameters for optimal performance, and evaluate its accuracy on a validation set using metrics

Data Collection

- Data Collection Process Overview:
- SpaceX REST API:
  - Collect data from SpaceX using their REST API.
- Data Wrangling:
  - Clean, preprocess, and handle missing values.
- One Hot Encoding:
  - Convert categorical variables for machine learning.
- Column Selection:
  - Remove irrelevant columns for focused analysis.
- Exploratory Data Analysis (EDA):
  - Analyze data distribution and relationships.
- Interactive Visual Analytics:
  - Utilize Folium and Plotly Dash for interactive visualizations.
- Classification Modeling:
  - Apply machine learning models for rocket landing prediction.
- Model Tuning and Evaluation:
  - Fine-tune models, evaluate using metrics, and iterate for optimization.

Data Collection – SpaceX API

- Getting Respone from API
- Creating Response to a .json file
- Apply Custom Functions to Clean Data
- Assing List to Dictionary then Dataframe
- Filter Dataframe and Export to Flat File (.csv)

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
response = requests.get(spacex_url).json()
```

```
data= pd.json_normalize(response)
```

```
getLaunchSite(data)
getPayloadData(data)
getCoreData(data)
getBoosterVersion(data)
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

```
df = pd.DataFrame.from_dict(launch_dict)
```

```
data_falcon9 = df.loc[df['BoosterVersion']!="Falcon 1"]
```

```
data_falcon9.to_csv('dataset_part_1.csv', index=False)
```

Data Collection - Scraping

- Getting Respone from HTML
- Creating BeautifulSoup Object
- Finding Tables
- Getting Column Names
- Creation of Dictionary
- Appending Data to keys
- Converting Dictionary to Dataframe
- Dataframe to .CSV

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-
```

```
page= requests.get(static_url)
```

```
soup= BeautifulSoup(page.text, 'html.parser')
```

```
html_tables= soup.find_all('table')
```

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

```
df = pd.DataFrame.from_dict(launch_dict)
```

```
df.to_csv('spaces_web_scraped.csv', index=False)
```

Data Wrangling

Data Cleaning:

- Handle missing values, outliers, and inconsistencies in the raw data.
- Utilize methods such as imputation or removal based on data characteristics.

One Hot Encoding:

- Convert categorical variables into a numerical format for machine learning.
- Use One Hot Encoding to represent categorical data as binary vectors.

Column Selection:

- Remove irrelevant columns that do not contribute to the analysis.
- Retain only essential features for streamlined datasets.

EDA with Data Visualization

Scatter Graphs Being Drawn:

- Flight Number vs. Payload Mass
- Flight Number vs. Launch Site
- Payload vs. Launch Site
- Orbit vs. Flight Number
- Payload vs. Orbit Type
- Orbit vs. Payloady Mass

Bar Graph Being Drawn:

- Mean vs. Orbit

Line Graph Being Drawn:

- Success Rate vs. Year

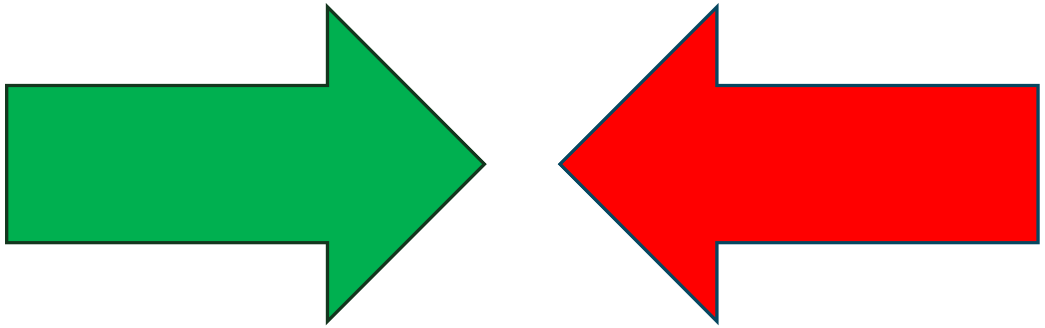
EDA with SQL

Which we are using SQL Queries to get the answers in the dataset:

- Displaying 5 records where launch sites begin with the string ‘KSC’
- Listing the date where the successful landing outcome in drone ship was achieved
- Listing total number of successful and failure mission outcomes
- Ranking the count of successful landing outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

Build an Interactive Map with Folium

We assigned the dataframe launch_outcomes(failures, successes) to classes 0 adn 1 with Green And Red markers on the map in a MarkerCluster.



Build a Dashboard with Plotly Dash

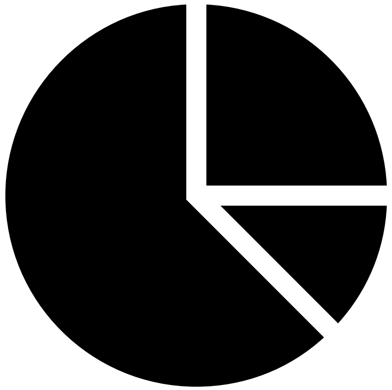
Graphs

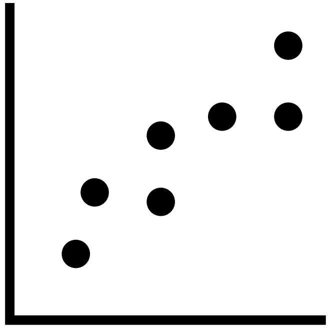
- Pie Chart

The total launches by a certain site/all sites

- Scatter Graph

The relationship between two variables





Predictive Analysis (Classification)

Model Selection:

- Choose classification algorithms suitable for the problem (e.g., logistic regression, decision trees, support vector machines).

Data Splitting:

- Divide the dataset into training and testing sets to assess model generalization.

Model Training:

- Train the chosen models on the training set to learn patterns and relationships.

Hyperparameter Tuning:

- Fine-tune model hyperparameters to optimize performance, utilizing techniques like grid search or randomized search.

Cross-Validation:

- Implement k-fold cross-validation to assess model robustness and avoid overfitting.

Evaluation Metrics:

- Assess model performance using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

Iterative Improvement:

- Iteratively refine models based on insights gained during the evaluation process.

Feature Engineering:

- Consider feature engineering techniques to enhance model interpretability and effectiveness.

Testing on Unseen Data:

- Evaluate the final model on the testing set to ensure it performs well on new, unseen data.

Documentation:

- Document the entire process, including model selection, hyperparameter values, and evaluation results.

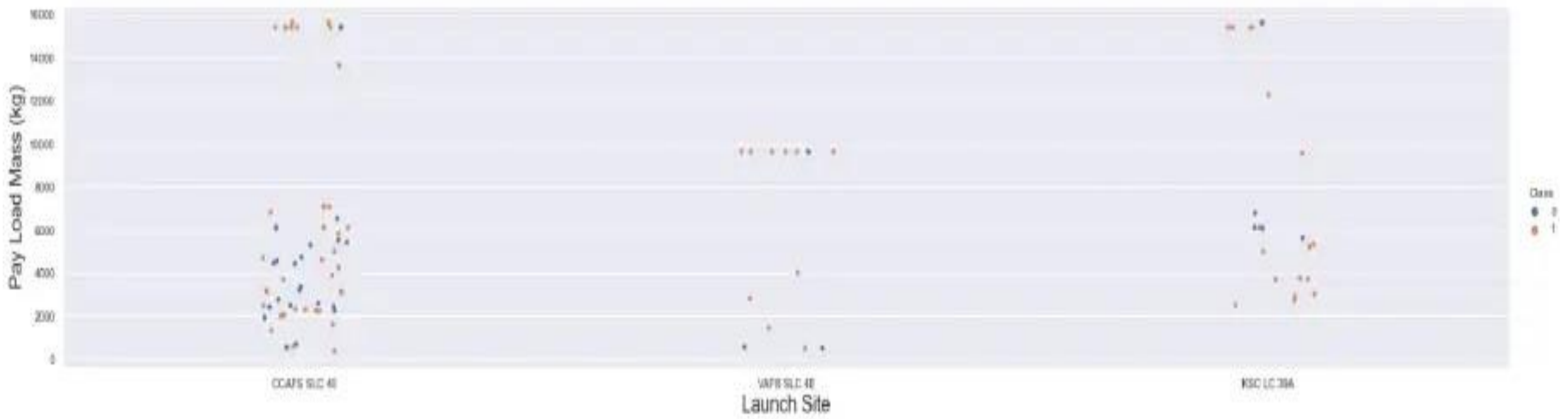
Results

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

Flight Number vs. Launch Site

The greater the number of flights at a launch site, the higher the success rate at that launch site.

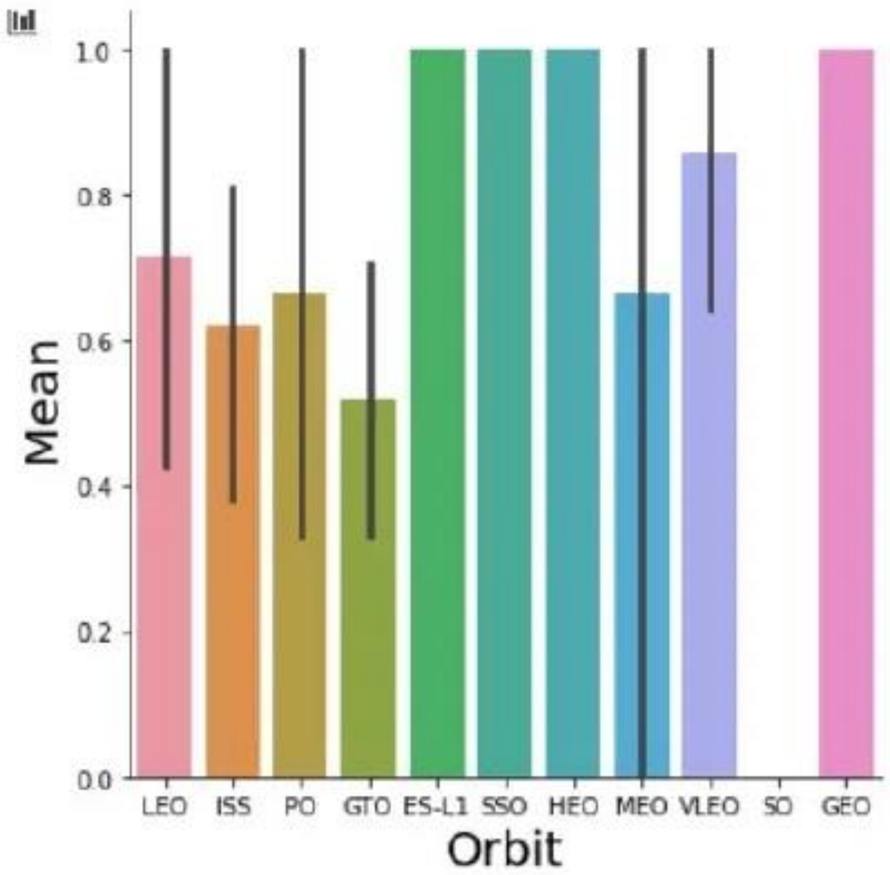
Payload vs. Launch Site



The higher the payload for the launch site CCAFS SLC-40, the greater the success rate for the rocket.

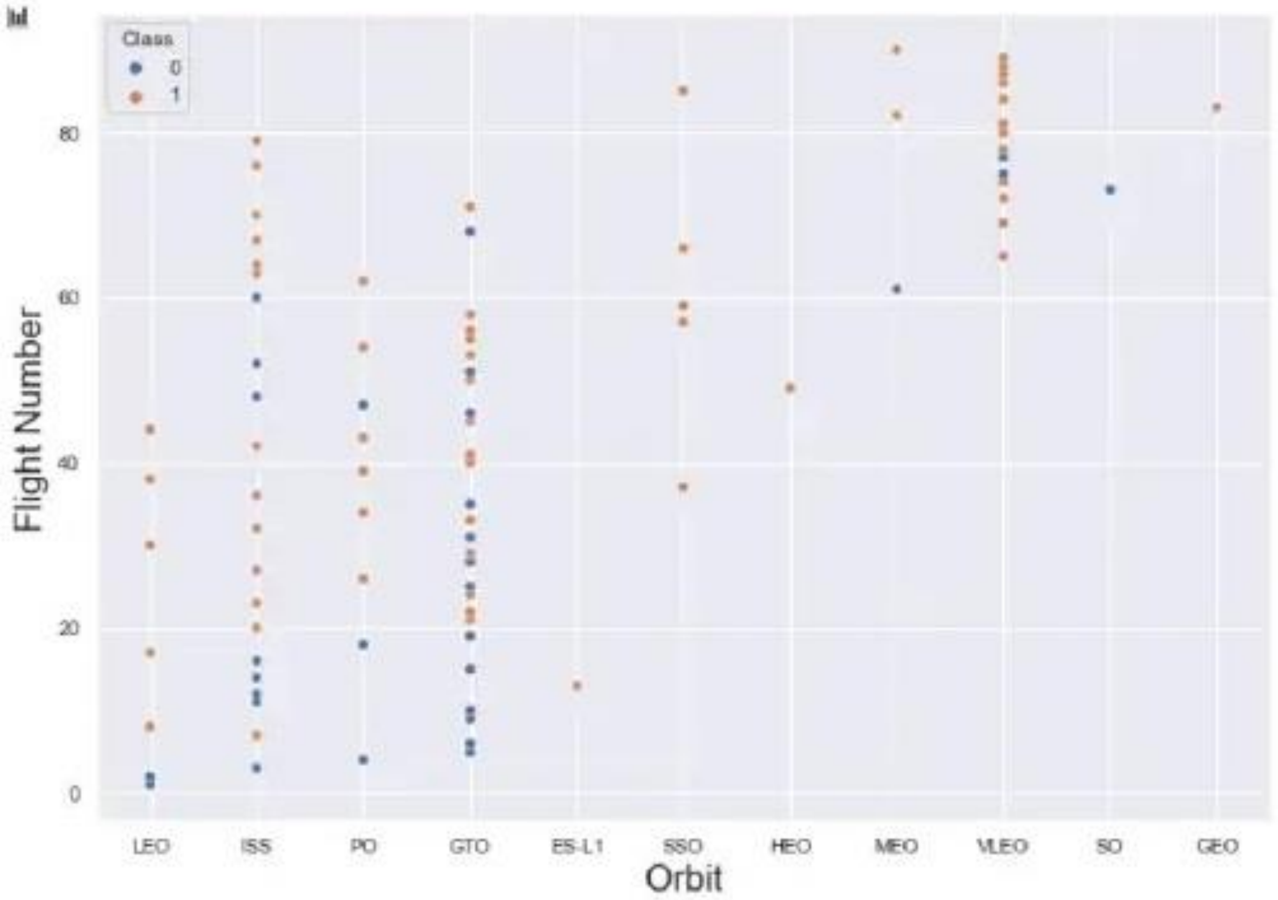
Success Rate vs. Orbit Type

Orbit GEO, HEO, SSO, ES-L1 has the best Success Rate.



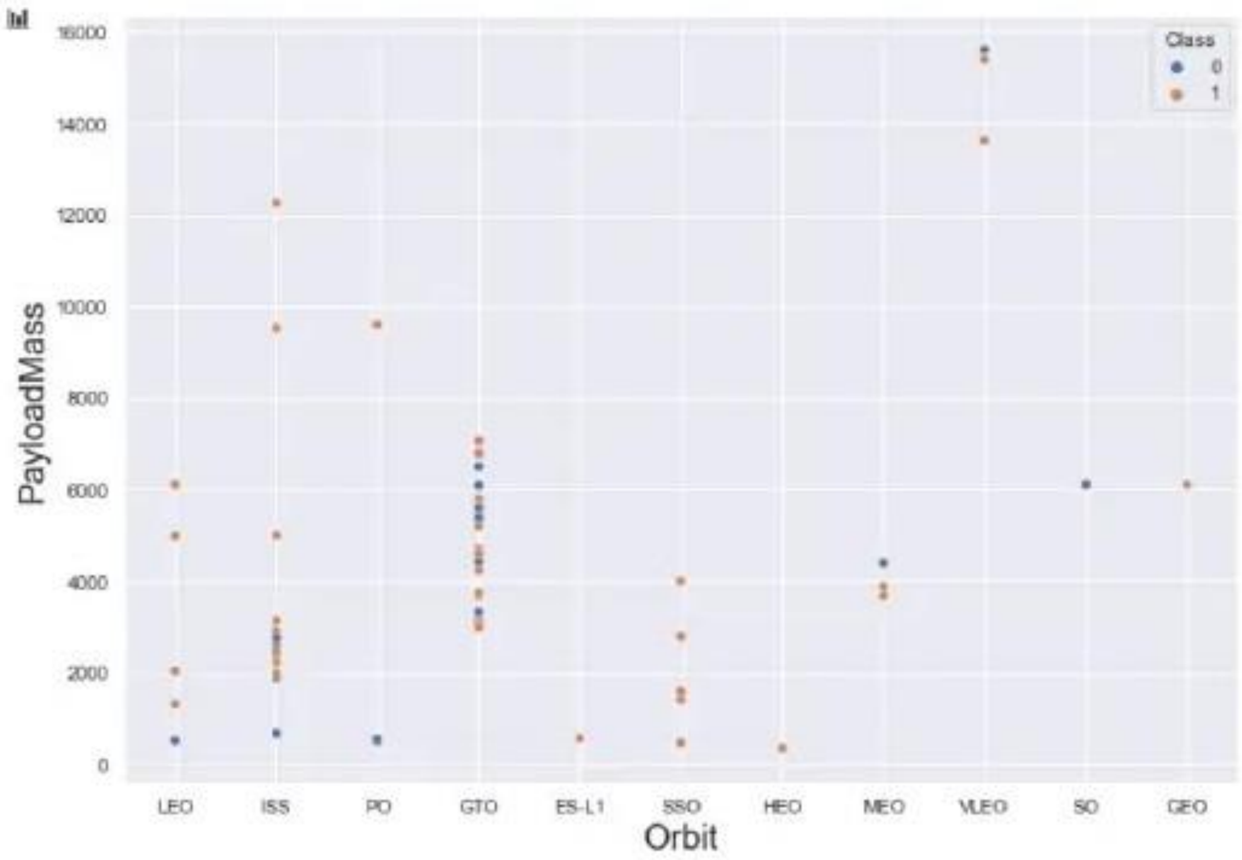
Flight Number vs. Orbit Type

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



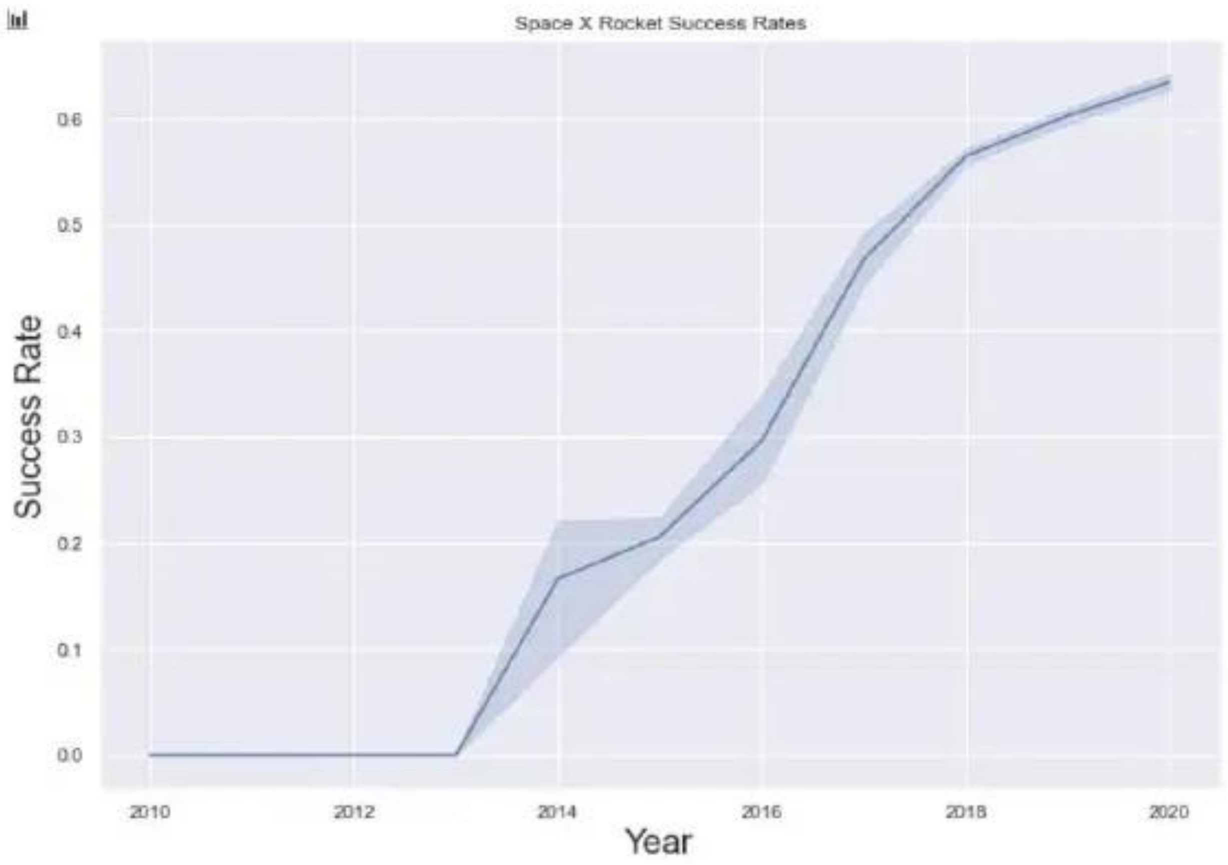
Payload vs. Orbit Type

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



Launch Success Yearly Trend

The success rate since 2013 kept increasing till 2020.



All Launch Site Names

```
SELECT DISTINCT Launch_Site
FROM tblSpaceX
```

UNIQUE LAUNCH SITES

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

Launch Site Names Begin with 'CCA'

SELECT TOP 5 *

FROM tblSpaceX

WHERE Launch_Site LIKE ‘KSC%’

	Date	Time_UTC	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS_KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
0	19-02-2017	2021-07-02 14:39:00.0000000	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (ground pad)
1	16-03-2017	2021-07-02 06:00:00.0000000	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No attempt
2	30-03-2017	2021-07-02 22:27:00.0000000	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (drone ship)
3	01-05-2017	2021-07-02 11:15:00.0000000	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (ground pad)
4	15-05-2017	2021-07-02 23:21:00.0000000	F9 FT B1034	KSC LC-39A	Inmarsat-5 F4	6070	GTO	Inmarsat	Success	No attempt

Total Payload Mass

SELECT SUM (PAYLOAD_MASS_KG) AS Total Payload Mass

FROM tblSpaceX

WHERE Customer = ‘NASA (CRS)’, TotalPayloadMass

	Total Payload Mass
0	45596

Average Payload Mass by F9 v1.1

SELECT AVG (PAYLOAD_MASS_KG) AS Average Payload Mass

FROM tblSpaceX

WHERE Booster_Version = ‘F9 v1.1’

	Average Payload Mass
0	2928

First Successful Ground Landing Date

```
SELECT MIN (Date)

FROM tblSpaceX

WHERE Landing_Outcome = ‘Success (drone ship)’
```

	Date which first Successful landing outcome in drone ship was acheived.
0	06-05-2016

Successful Drone Ship Landing with Payload between 4000 and 6000

```
SELECT Booster_Version
FROM tblSpaceX
WHERE Landing_Outcome = ‘Success (ground pad)’
AND Payload_Mass_KG > 4000
AND Payload_Mass_KG < 6000
```

	Date which first Successful landing outcome in drone ship was acheived.
0	F9 FT B1032.1
1	F9 B4 B1040.1
2	F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

```
SELECT (SELECT COUNT (Mission_Outcome)
FROM tblSpaceX
WHERE (Mission_Outcome LIKE ‘%SUCCESS%’) AS Successful_Mission_Outcomes,
(SELECT COUNT (Mission_Outcome)
FROM tblSpaceX
WHERE Mission_Outcome LIKE ‘%Failure%’ AS Successful_Mission_Outcomes)
```

	Successful_Mission_Outcomes	Failure_Mission_Outcomes
0	100	1

Boosters Carried Maximum Payload

SELECT DISTINCT Booster_Version, MAX(Payload_Mass_Kg) AS Maximum Payload Mass

FROM tblSpaceX

GROUP BY Booster_Version

ORDER BY Maximum Payload Mass DESC

	Booster_Version	Maximum Payload Mass
0	F9 B5 B1048.4	15600
1	F9 B5 B1048.5	15600
2	F9 B5 B1049.4	15600
3	F9 B5 B1049.5	15600
4	F9 B5 B1049.7	15600
...	...	...
92	F9 v1.1 B1003	500
93	F9 FT B1038.1	475
94	F9 B4 B1045.1	362
95	F9 v1.0 B0003	0
96	F9 v1.0 B0004	0

97 rows × 2 columns

2015 Launch Records

```
SELECT DATENAME (month,
DATEADD(month,

MONTH(CONVERT(date, Date, 105)),0) -1) AS
Month, Booster_Version, LAUNCH_Site,
Landing_Outcome

FROM tblSpaceX

WHERE (Landing_Outcome LIKE
'%SUCCESS%')

AND  (YEAR(CONVERT(date, Date, 105)) =
'2015')
```

Month	Booster_Version	Launch_Site	Landing_Outcome
January	F9 FT B1029.1	VAFB SLC-4E	Success (drone ship)
February	F9 FT B1031.1	KSC LC-39A	Success (ground pad)
March	F9 FT B1021.2	KSC LC-39A	Success (drone ship)
May	F9 FT B1032.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1035.1	KSC LC-39A	Success (ground pad)
June	F9 FT B1029.2	KSC LC-39A	Success (drone ship)
June	F9 FT B1036.1	VAFB SLC-4E	Success (drone ship)
August	F9 B4 B1039.1	KSC LC-39A	Success (ground pad)
August	F9 FT B1038.1	VAFB SLC-4E	Success (drone ship)
September	F9 B4 B1040.1	KSC LC-39A	Success (ground pad)
October	F9 B4 B1041.1	VAFB SLC-4E	Success (drone ship)
October	F9 FT B1031.2	KSC LC-39A	Success (drone ship)
October	F9 B4 B1042.1	KSC LC-39A	Success (drone ship)
December	F9 FT B1035.2	CCAFS SLC-40	Success (ground pad)

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

```
SELECT COUNT ( Landing_Outcome)

FROM tblSpaceX

WHERE (Landing_Outcome LIKE ‘%SUCCESS%’)

AND (Date > ’04-06-2010’)

AND (Date > ’20-03-2017’)
```

	Successful Landing Outcomes Between 2010-06-04 and 2017-03-20
0	34

Folium Map – All Launch Sites Global Map Markers

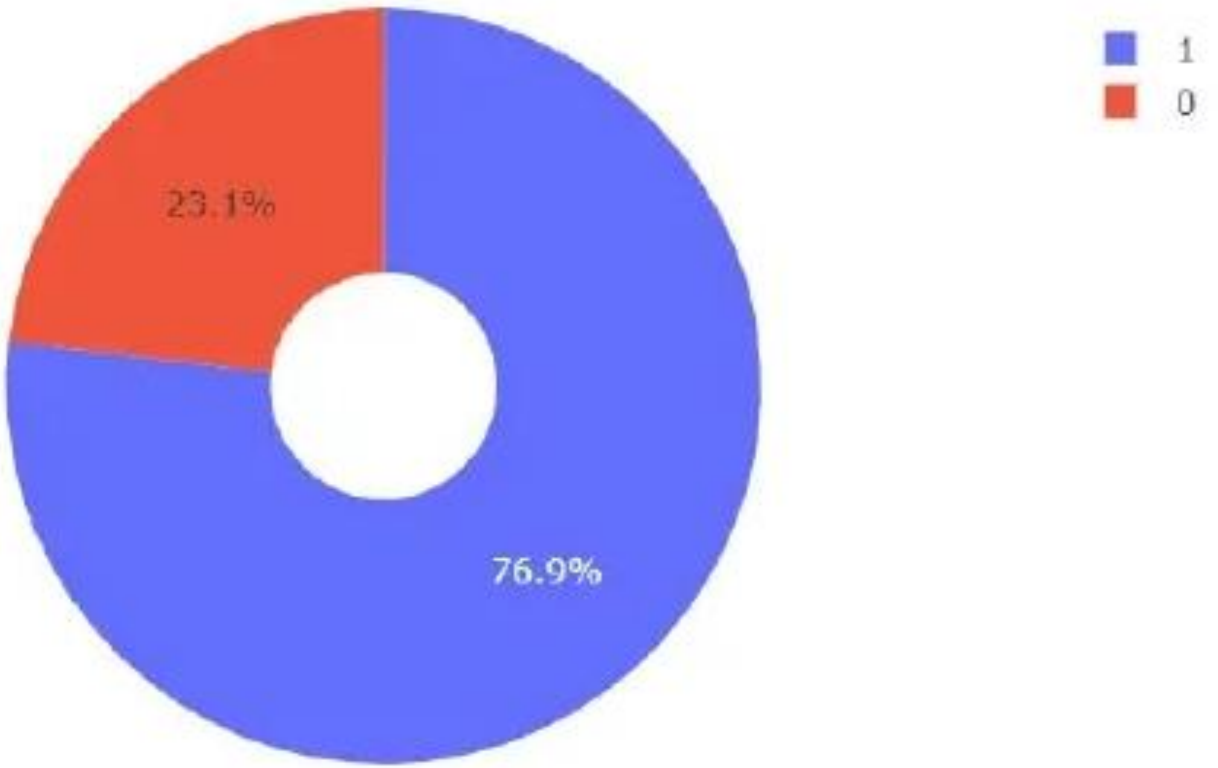
Folium Map – Colour Labelled Markers

Folium Map – Haversine Formula

Dashboard with Plotly Dash

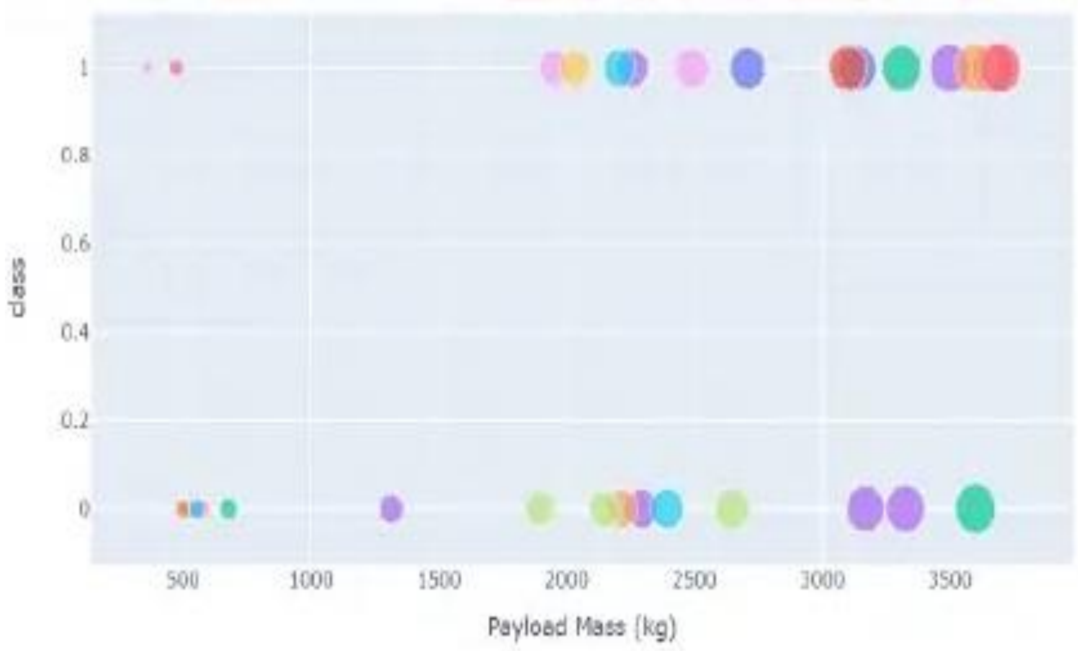
Dashboard with Plotly Dash

KSC LC-39A achieved a 76.9% success rate.

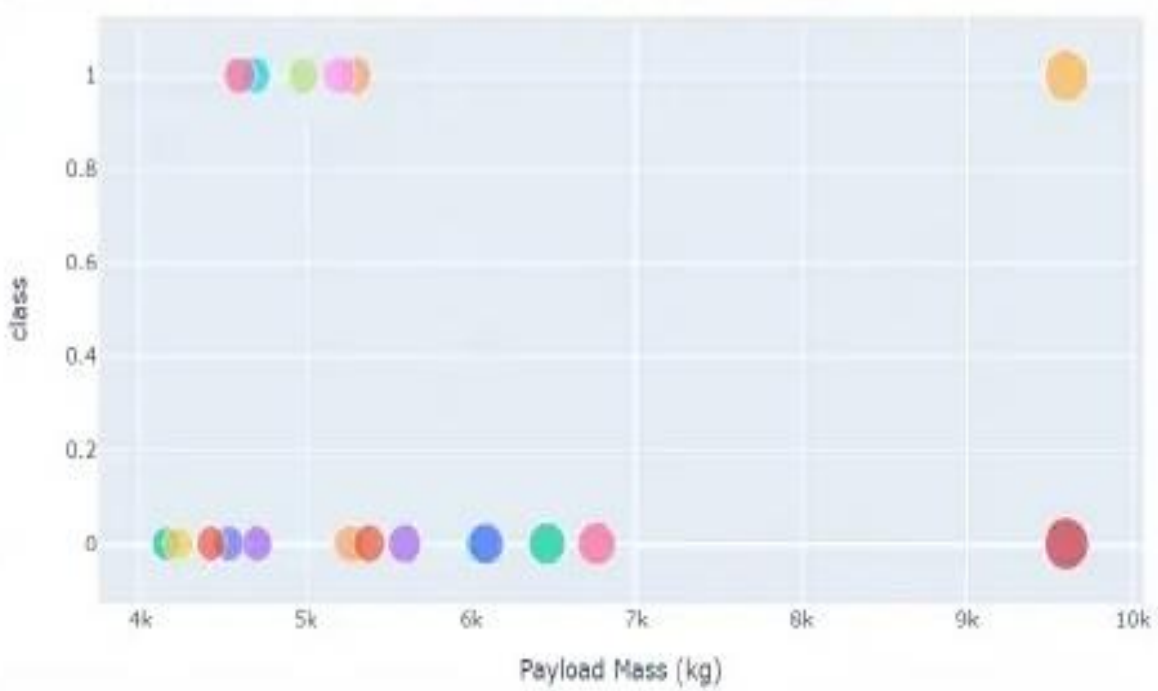


Dashboard with Plotly Dash

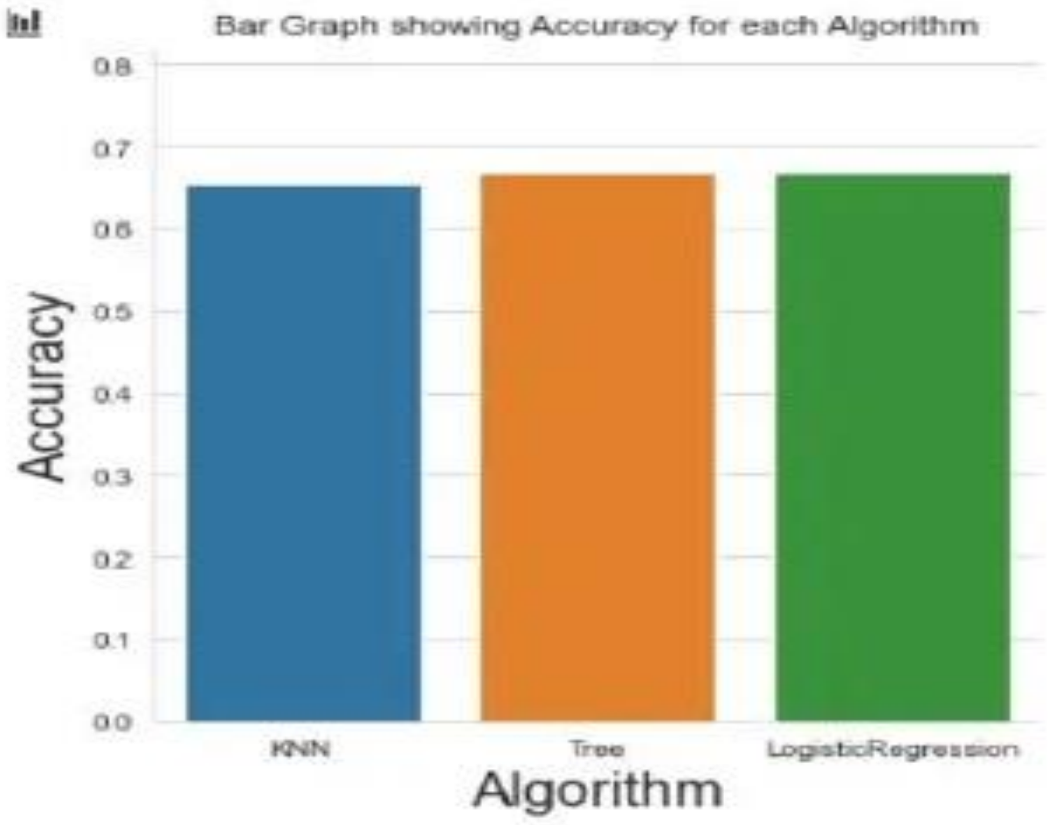
Low Weighted PAYLOAD 0 KG – 4000 KG



Heavy Weighted PAYLOAD 0 KG – 4000 KG



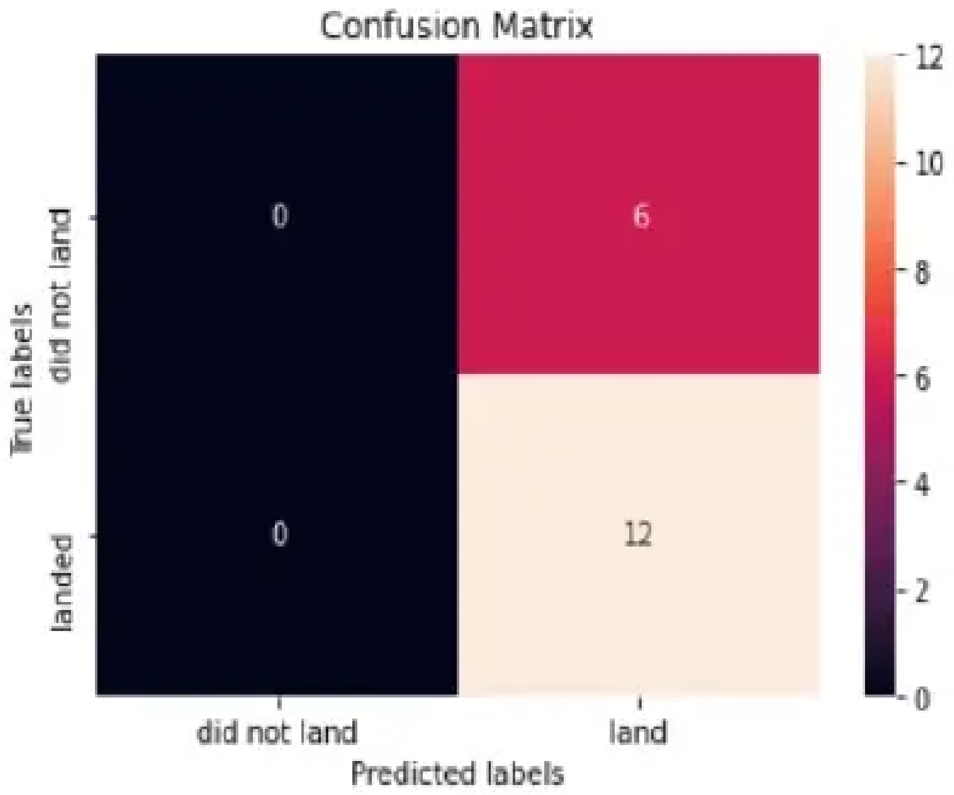
Classification Accuracy



Our accuracy is extremely close !

Confusion Matrix

Examining the confusion matrix, we see that Tree can distinguish between the different classes. We see that the major problem is false positives.



Conclusions

- The Tree Classifier Algorithm is the best for MACHİNE Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- We can see that KSC LC-39A had the most successful launches from all the sites
- Orbit GEO, HEO, SSO, ES-L1 has the best Success Rate

Appendix

- Haversine Formula
- ADGGoogleMaps Module
- Module SQL Server
- Python 3 Jupiter