Winning Space Race with Data Science

DEVAM SINGH

13-06-24

MY GITHUB- https://github.com/Devamsingh09

Outline

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

Executive Summary

- Methodology Summary

We performed CRISP-DM (Cross-Industry Standard Process for Data Mining) Methodology
Here's how the CRISP-DM methodology is generally applied :

1. ***Business Understanding***
2. ***Data Understanding***
3. ***Data Preparation***
4. ***Modeling***
5. ***Evaluation***
6. ***Deployement***
7. ***Feedback***

Introduction

- Project background and context

In this report, we predicted if the Falcon 9 first stage will land successfully. SpaceX 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. 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 sources we used to gather information about SpaceX, and how did we clean and prepared this data for analysis?
•What are the key metrics and variables we collected about SpaceX's missions and first-stage landings?
•Can we provide a summary of the data, including the number of missions, success rates, and any trends or patterns observed?

Introduction

- Problems you want to find answers

•Describe the process of building and training the machine learning model to predict if SpaceX will reuse the first stage.

•Can you explain the choice of the machine learning algorithm(s) used for this prediction task?
•How did you evaluate the performance of the model, and what metrics did you use to assess its accuracy?

•How confident is the model in its predictions, and what are the implications of these predictions for SpaceX's operations?

•Based on the model's insights, what recommendations would you provide to SpaceX regarding their first-stage reuse strategy?

•Were there any limitations or challenges faced during the model-building process, and how were they addressed?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:

  The capstone assignment involves working with SpaceX launch data obtained from the SpaceX REST API. This API provides information about launches, including details about the rocket used, payload delivered, launch specifications, landing specifications, and a summary of the landing outcome.

- Perform data wrangling

  The dataset contains Flight Number, Date, Booster version, Payload mass, Orbit, Launch Site, Outcome (landing status), Grid Fins, Reused, Legs, Landing pad, Block, Reused count, Serial, and Longitude/Latitude. Launch sites: Vandenberg AFB, Kennedy Space Center, CCAFS SLC 40. Orbits: LEO, GTO. "Outcome" shows successful ("True ASDS") or unsuccessful ("False ASDS") first-stage landings, to be converted to binary (0 for unsuccessful, 1 for successful) for "Y."

Methodology

- Perform exploratory data analysis (EDA) using visualization and SQL

  In the project, will determine what attributes are correlated with successful landings.
  The categorical variables will be converted using one hot encoding, preparing the data for a machine learning model that will predict if the first stage will successfully land.

- Perform interactive visual analytics using Folium

  With interactive visual analytics, users could find visual patterns faster and more effectively.
  Instead of presenting our findings in static graphs, interactive data visualization, or dashboarding, can always tell a more appealing story.
  In this project, will be using Folium.

- Perform predictive analysis using classification models

  We will test Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors.
  Finally, we will output the confusion matrix.

Data Collection

- Describe how data sets were collected.

In this analysis, we will review Collecting the Data with an API.

In this assignment, we will be working with SpaceX launch data that is gathered from an API, specifically the SpaceX REST API.

This API will give us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.

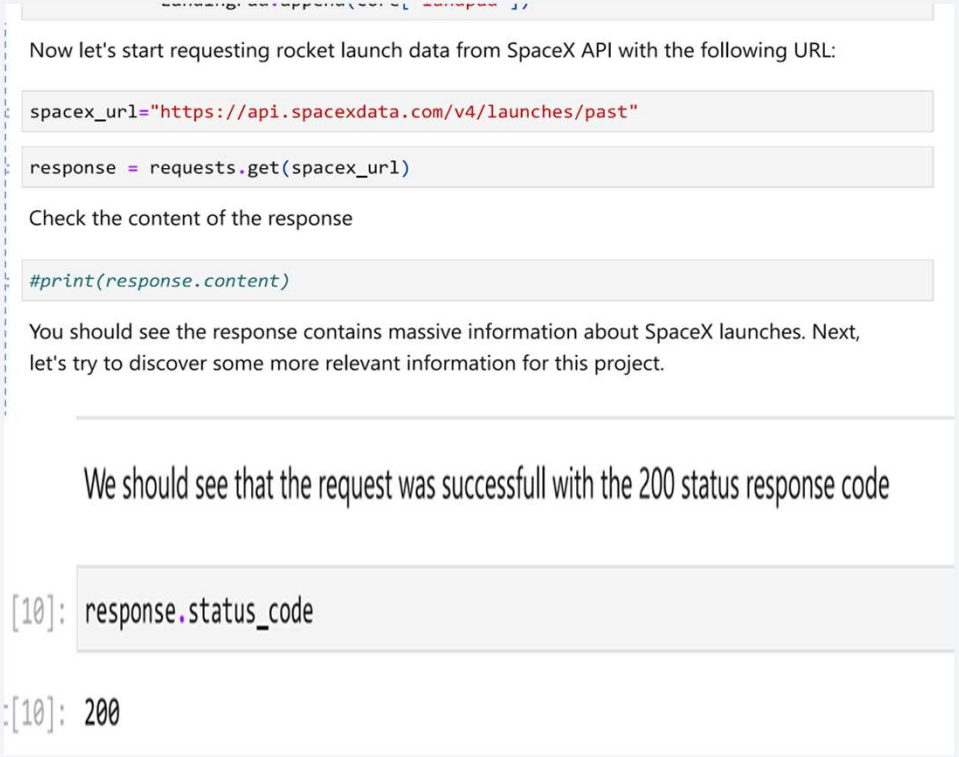
Our goal is to use this data to predict whether SpaceX will attempt to land a rocket or not.

The SpaceX REST API endpoints, or URL, starts with api.spacexdata.com/v4/.

Data Collection – SpaceX API

- Using spacex URL we are putting it in spacex_url variable and generating the response through requests.get method.

- We can see API is working properly as resulted response code i.e. 200.
- API_URL: https://api.spacexdata.com/v4/launches/past



Data Collection - Scraping

- After generating response we used BeautifulSoup for webscrapping. Here we extracted various tables and their <Th> elements .
- We created empty dictionaries with keys from extracted columns and made new dataframe.
- WebScrapping:https://github.com/Devamsingh09/SPACE_X_FALCON9.github.io/blob/main/2_SPACE_Xwebscrapping.ipynb

Next, we want to collect all relevant column names from the HTML table header

Let's try to find all tables on the wiki page first. If you need to refresh your memory about `BeautifulSoup`, please check the external reference link towards the end of this lab

```
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')

# Print the number of tables found
print("Number of tables found:", len(html_tables))
```

```
Number of tables found: 25
```

Data Wrangling

- How data were processed:

We would like landing outcomes to be converted to Classes y. y. (either 0 or 1). 0 is a bad outcome, that is, the booster did not land. 1 is a good outcome, that is, the booster did land. The variable Y will represent the classification variable that represents the outcome of each launch.

Calculating Num of Launches of each site:

```
# Apply value_counts() on column LaunchSite
df['LaunchSite'].value_counts()
```

```
CCAFS SLC 40     55
KSC LC 39A       22
VAFB SLC 4E      13
Name: LaunchSite, dtype: int64
```

Using the outcome we created a list where the

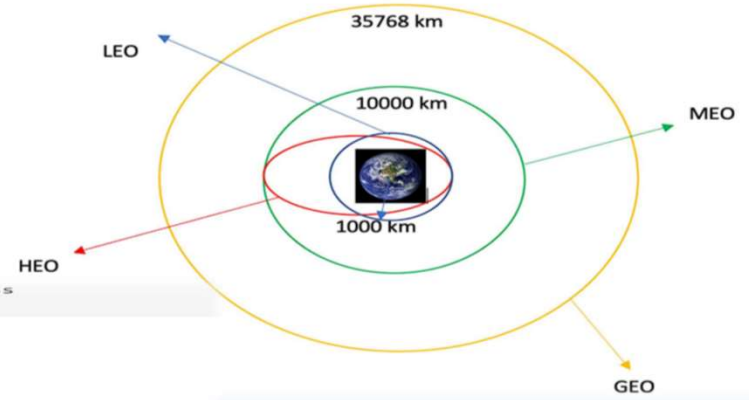
Element is 0 if the corresponding row in outcome is in the

Set bad outcome (i.e. its class value is 0)

```
df['Class']=landing_class
df[['Class']].head(8)
```

	Class
0	0
1	0
2	0
3	0
4	0
5	0
6	1

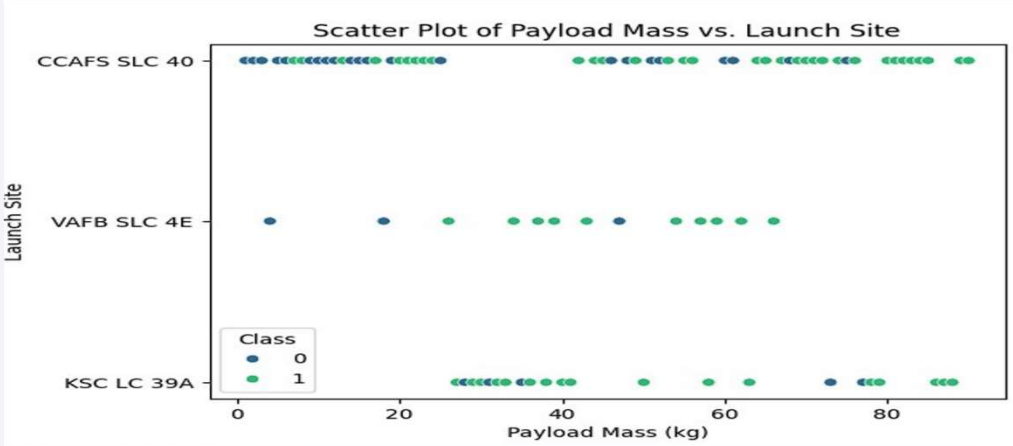
Here are some common dedicated orbits

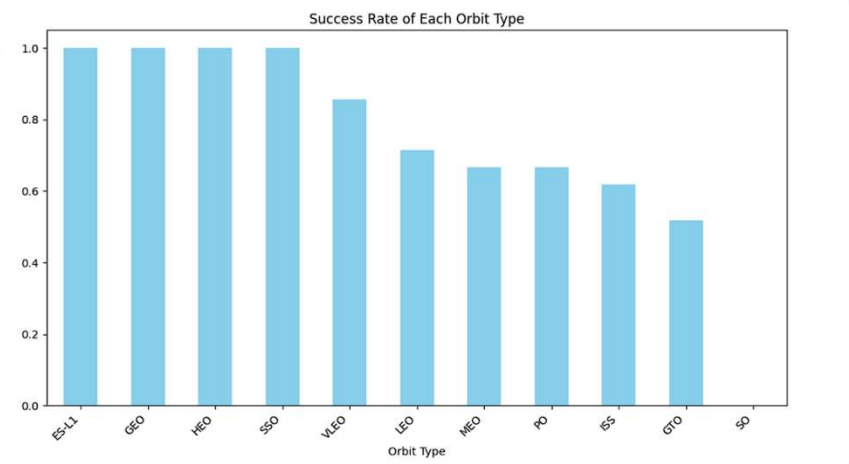


DataWrangling: https://github.com/Devamsingh09/SPACE_X_FALCON9.github.io/blob/main/3_Spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- We used mainly cat plot and scatterplot to find out the relationships between Payload Mass, Flight Num, Launch Site. We also plot bar chart with orbit success rate of all the orbits to find the most suitable orbit for future purpose.





EDA GITHUB Link:
https://github.com/Devamsingh09/SPACE_X_FALCON9.github.io/blob/main/5_Exploratory_Data_Analysis.ipynb

EDA with SQL

- We selected distinct Launch Site from SpaceX Table
- Selected Data Whose Launch Site Starts with something CCA i.e. CCAFS LC-40 and CCAFS SLC-40
- Selected AVG Pay Load Mass from Table where booster version is F9 V1.1.
- Selected first successful landing date where landing outcome is TRUE RTLS.
- Selected first successful landing date where landing outcome is TRUE RTLS and Pay Load > 4000.
- Selected list of cases where launch is successful assigning it as 1 else 0.
- Here are the Outcomes of the analysis:
- For more:

SQL EDA GITHUB Link:

https://github.com/Devamsingl
blob/main/4_SQL_statements.i

Out[37]:

Landing_Outcome	OutcomeCount
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

Build an Interactive Map with Folium

- We added folium circle and folium map marker to the all the launch sites of data to understand the launching sites quickly.
- We added marker cluster for better visualization and also coordinates for nearest highway, city and railway along with distance for convenience of readers.

- For more:

Folium GITHUB Link:

https://github.com/Devamsi
main/6_Mapping_and_locat

Predictive Analysis (Classification)

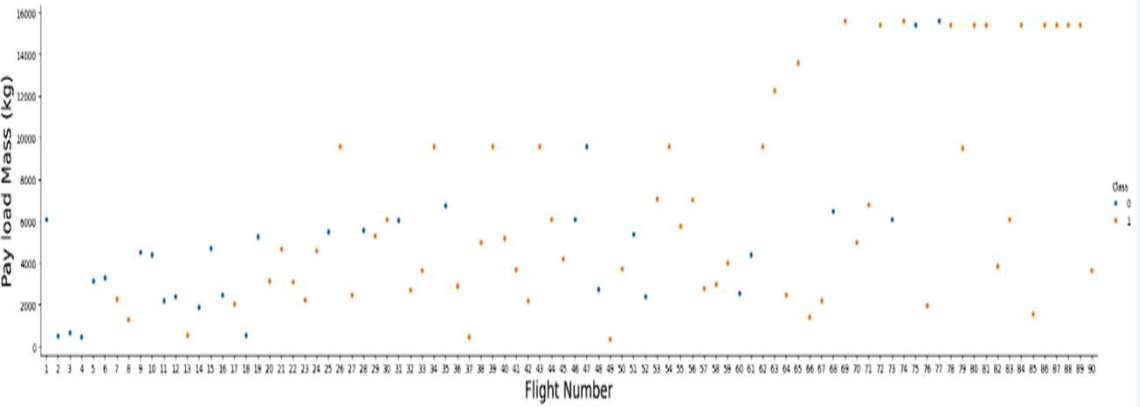
- ***Firstly, we used logistic regression and created Grid SearchCV object we fit logistic regression in GridSearch along with training dataset in which our target is 'Class' that is successfully launched or not. Got the accuracy of 83%.***
- ***Then we created a support vector machine and grid search object to find best parameters. This time we got 81% accuracy.***
- ***We also created confusion matrix to understand the prediction accuracy of model.***
- ***After we created decision tree classifier which provided the accuracy of 88%.***
- ***Then we created KNN Classifier and Grid Search for best parameters and got 61% accuracy.***
- ***Hence, we found our decision tree accuracy as best which is about 88%.***
- ***For more findings and outputs***:

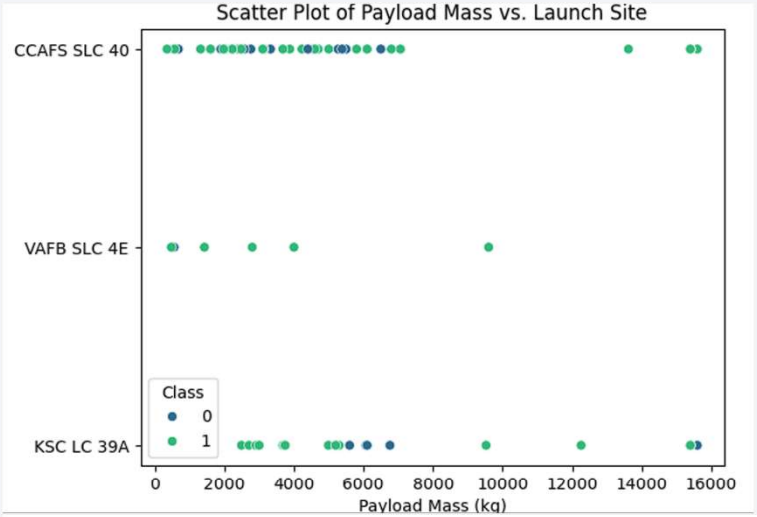
ML ALGORITHMS GITHUB Link:
https://github.com/Devamsingh09/SPACE_X_FALCON9.github.io/blob/main/7_Final_Part_SPACE_X_FALCON9_.ipynb

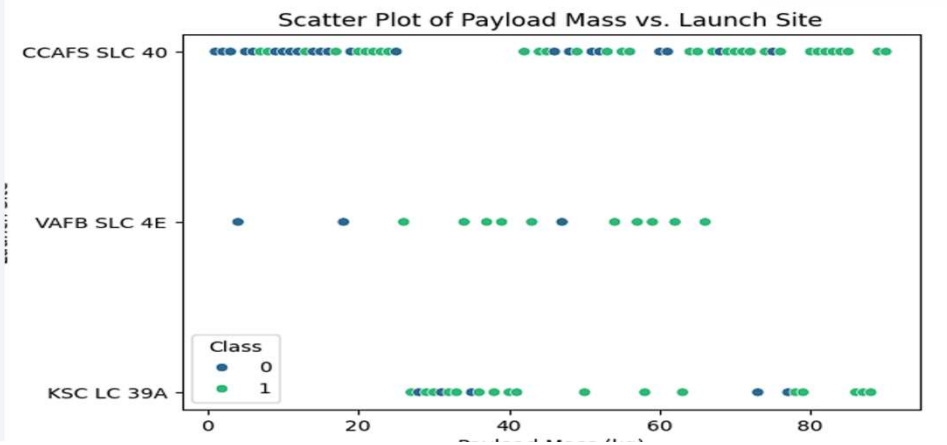
Results

- Exploratory data analysis results

```
sns.catplot(y="PayloadMass", x="FlightNumber", hue="Class", data=df, aspect = 5)
plt.xlabel("Flight Number",fontsize=20)
plt.ylabel("Pay load Mass (kg)",fontsize=20)
plt.show()
```



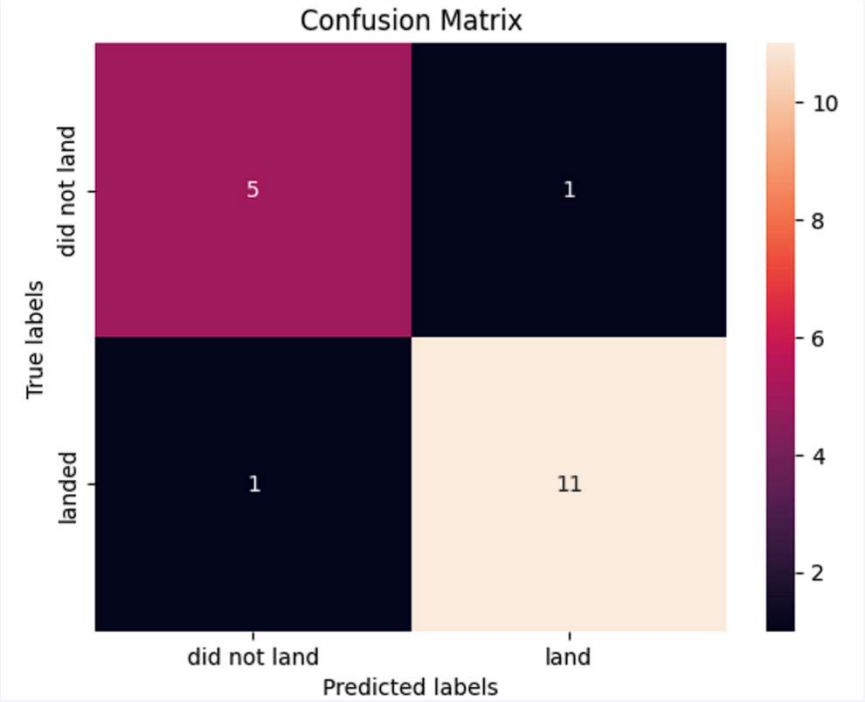




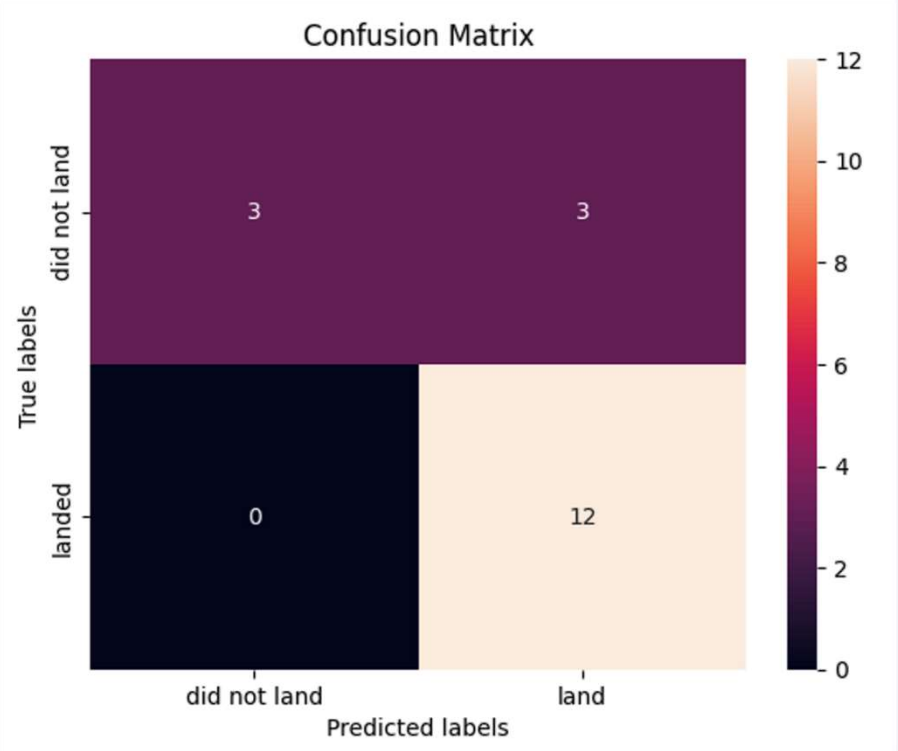
Results

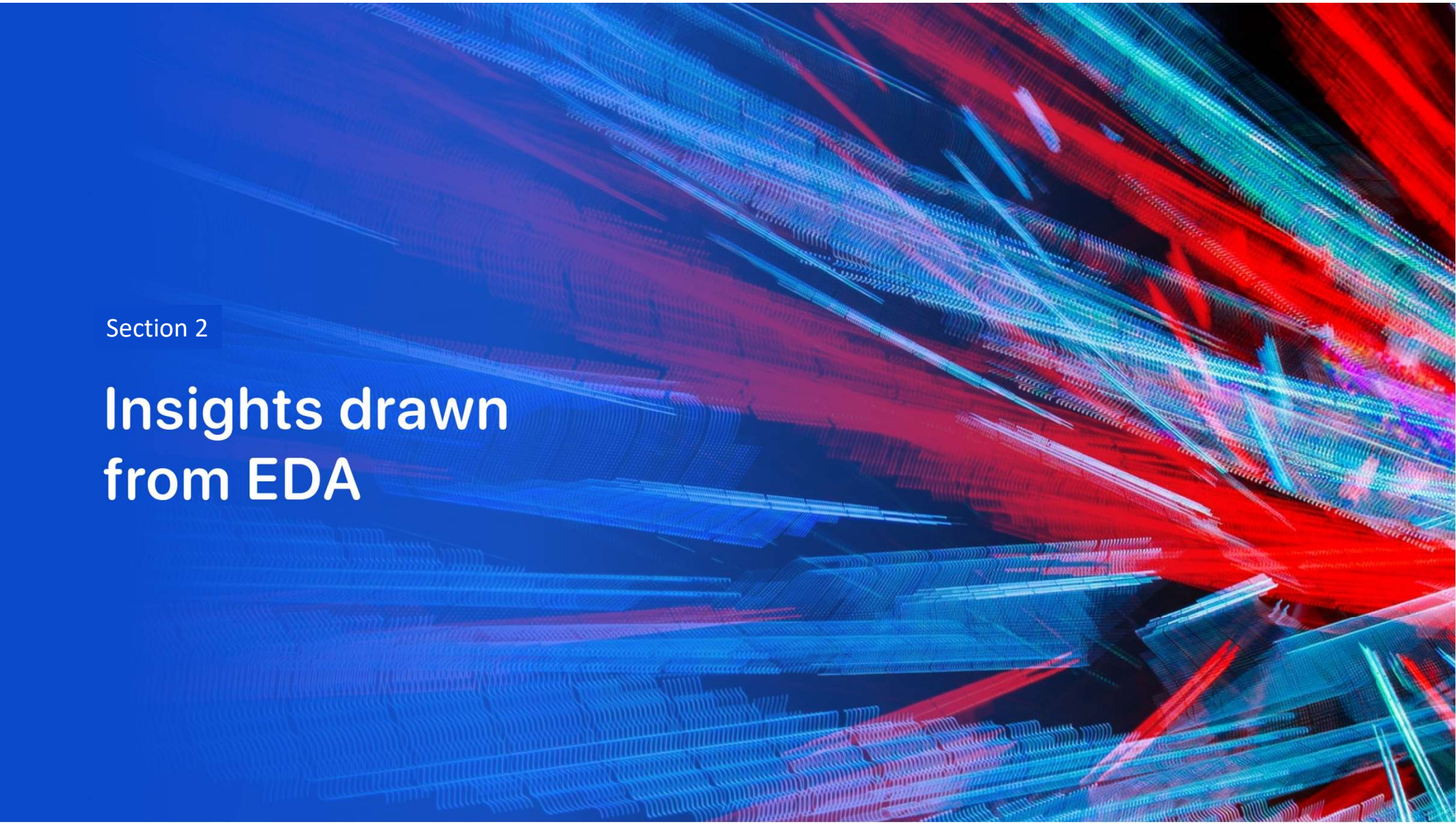
- **Predictive data analysis results**

For Classification Tree(88% accuracy)



For logistic regression



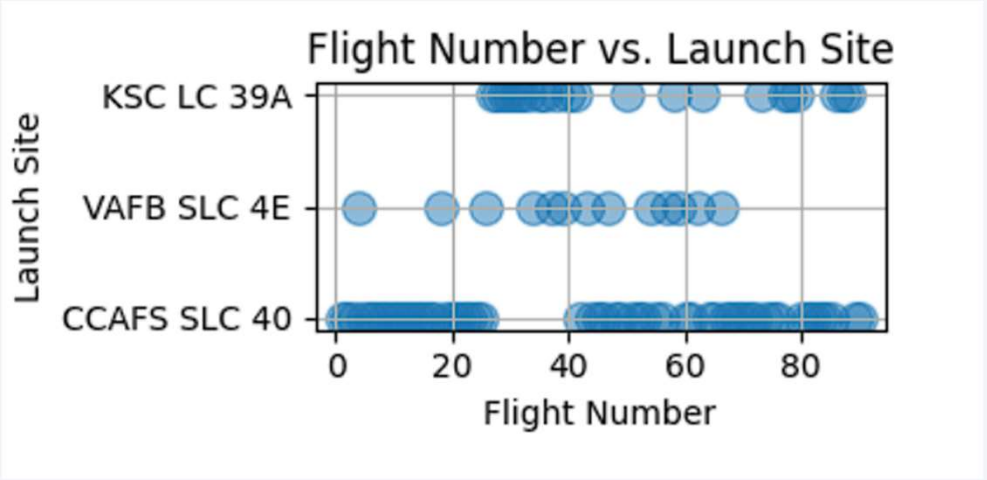


Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

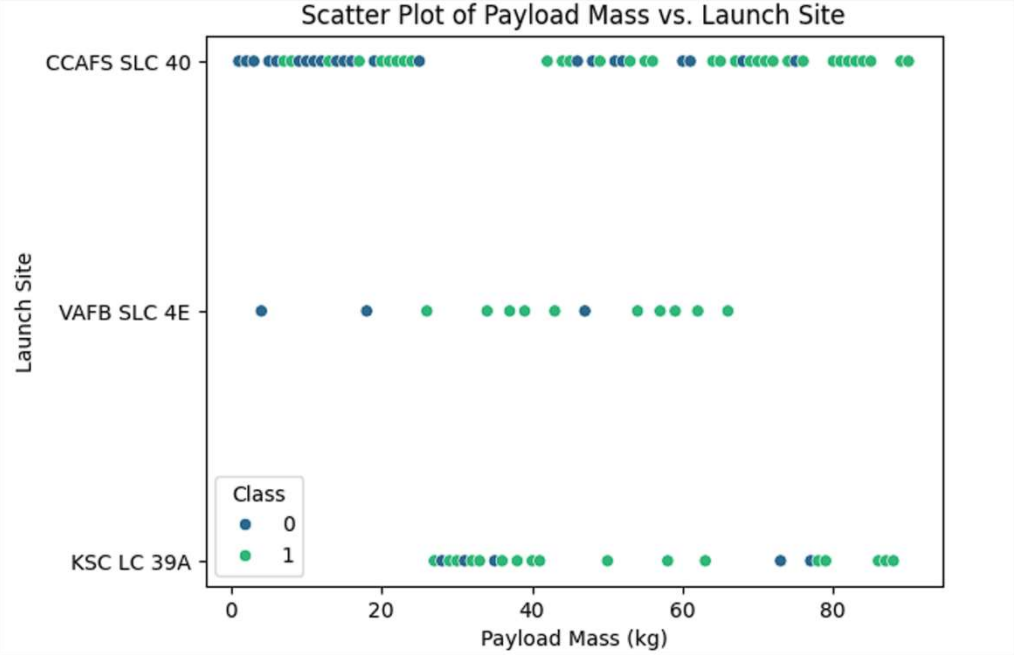
- Scatter plot of Flight Number vs. Launch Site



Explanation- Here number of flights occurred continuously in CCAFS SLC 40 site from starting and it contains the major part of course. The least number of flights occurred in launching site of VAFB SLC 4E.

Payload vs. Launch Site

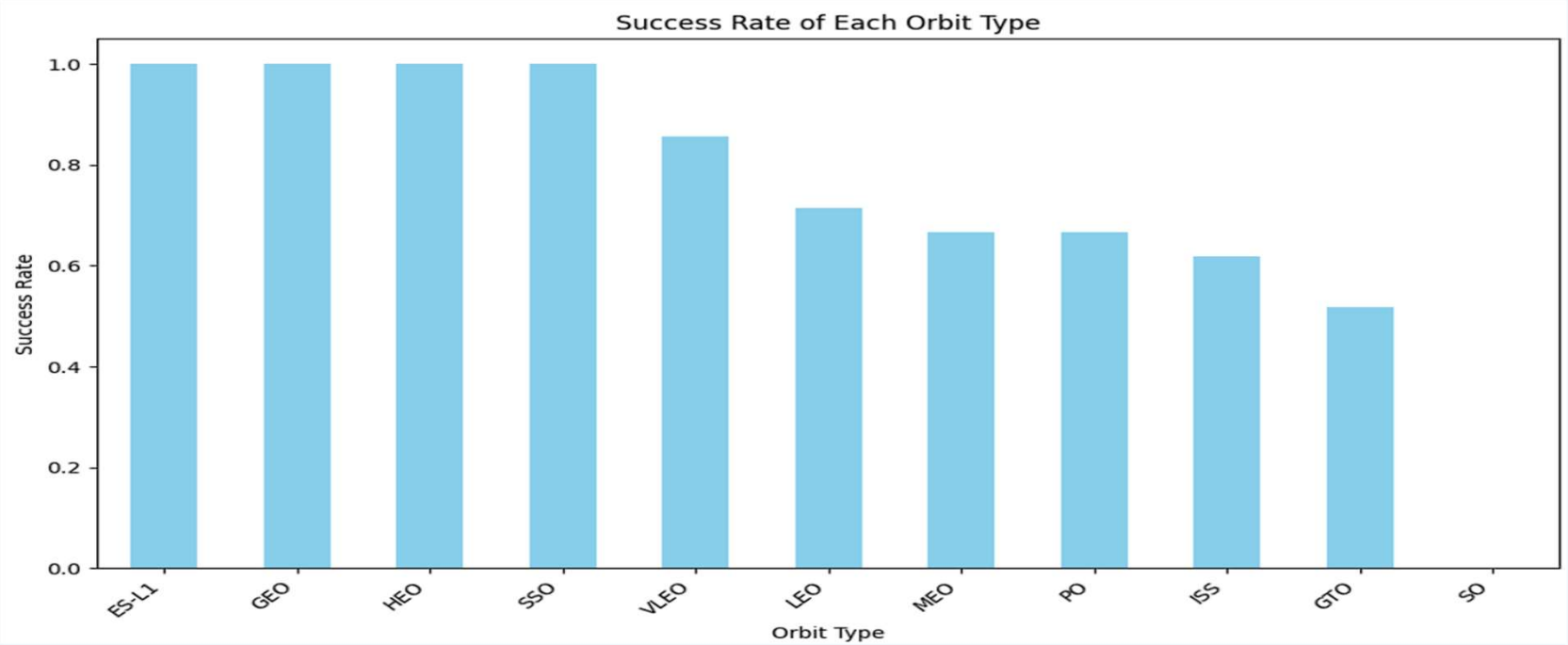
Scatter plot of Payload vs. Launch Site



Explanation- Here class 1 denotes successful launch and 0 denotes failed launch. We observe that least payload mass was flied from the launching site of VAFB SLC 4E and also very less in numbers but have high success rate. While highest payload was carried out from CCAFS SLC40 with low success rate for lower mases and higher for the higher payload masses.

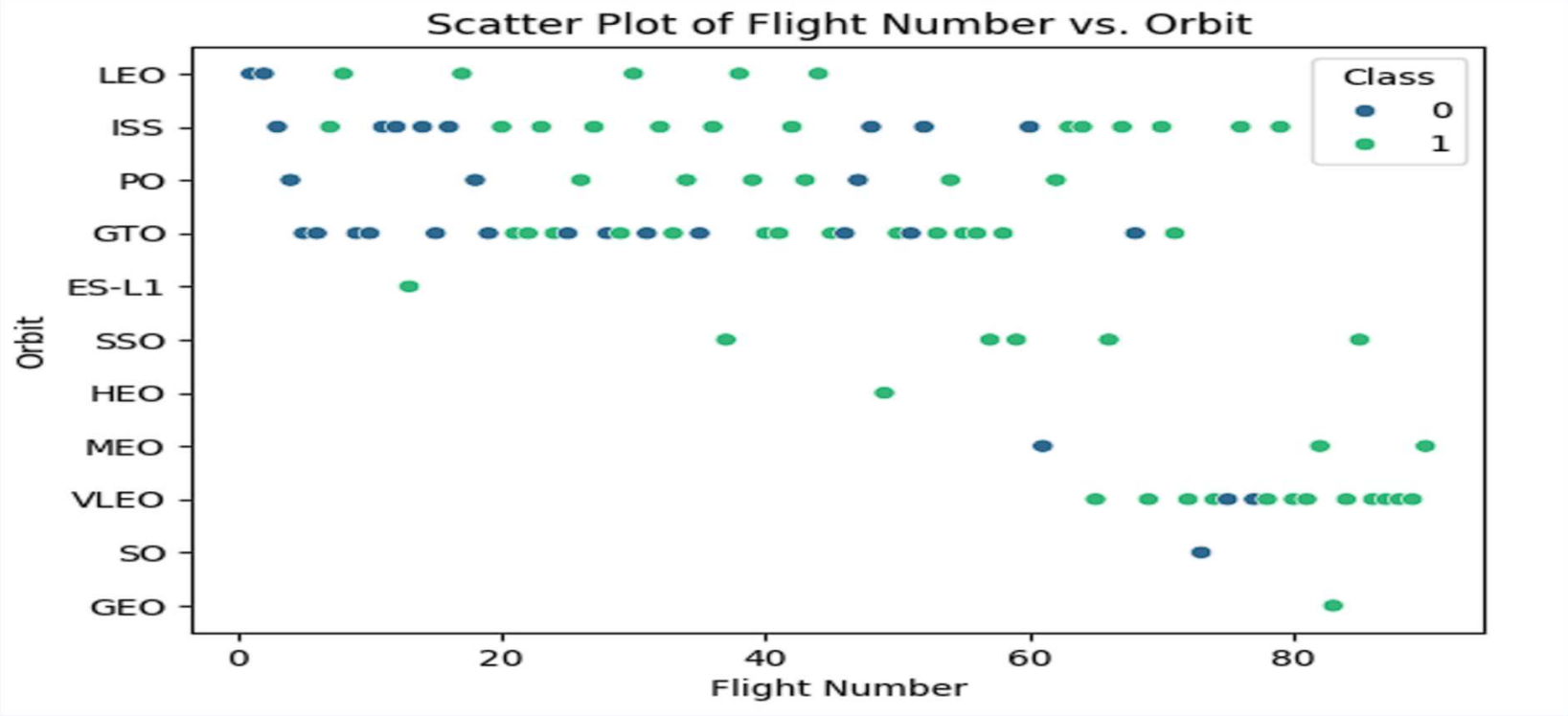
Success Rate vs. Orbit Type

Bar chart for the success rate of each orbit type



Explanation- Least success rate for SO and GTO orbits and best for ES-L1,GEO,HEO,SSO. We can consider these orbits for future rocket launchings.

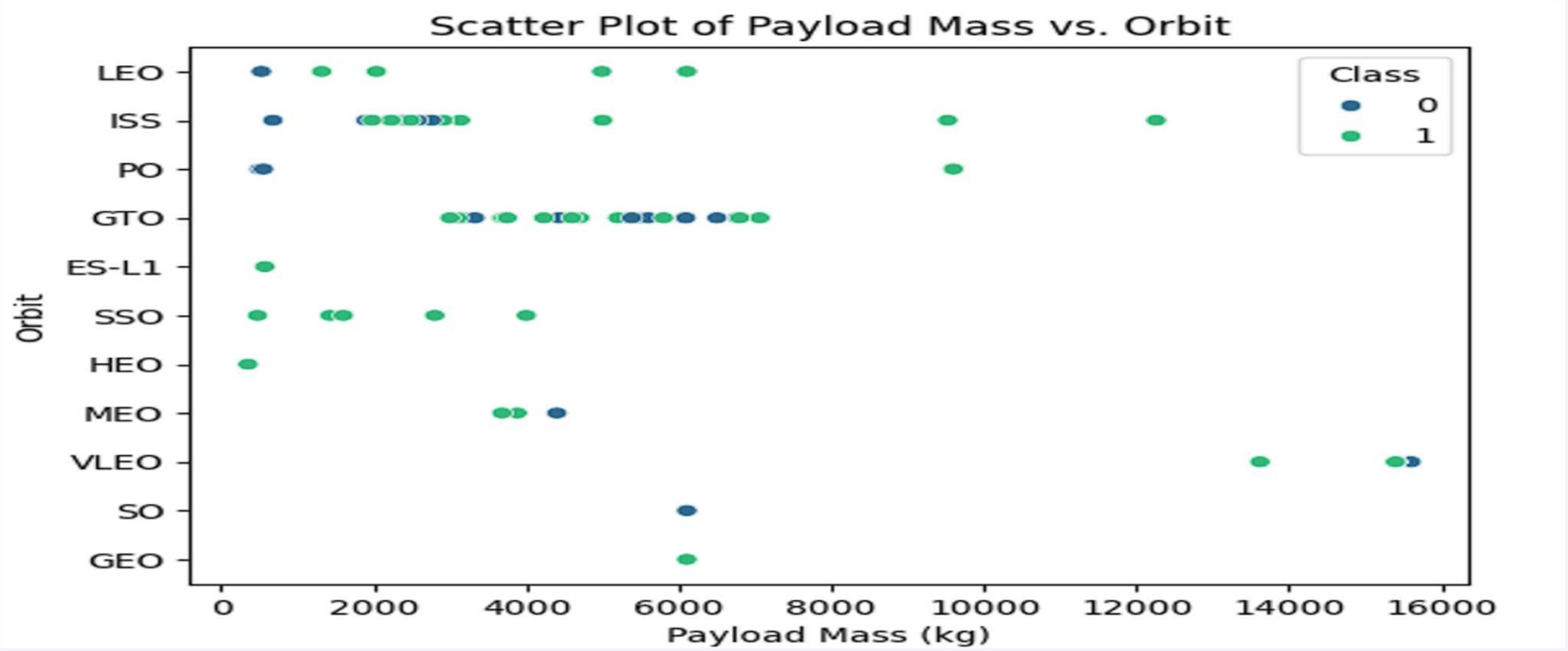
Flight Number vs. Orbit Type



Explanation- We performed only one flight in the orbits of GEO and SO with success and failure respectively. We can observe that only one failure in VLEO orbit and we can consider it for future launchings.

Payload vs. Orbit Type

Scatter plot of payload vs. orbit type



All Launch Site Names

- Names of the unique launch sites:

Here we used sql magic by %sql command of line and selected the unique Launch_Sites from SPACEXTABLE.
And the result is shown in the picture

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

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

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

%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Launch Site Names Begin with 'CCA'

5 records where launch sites begin with `CCA':

We used the query:

%sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5

Done.

Out[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

Total payload carried by boosters from NASA

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTABLE WHERE Customer = 'NASA (CRS)';
```

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

TOTAL_PAYLOAD_MASS
45596

Explanation- The above picture shows the SQL input query with result for the total Pay Load Mass when the customer was NASA (CRS). We selected the SUM function to find the total sum of payload mass.

Average Payload Mass by F9 v1.1

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

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AveragePayloadMass FROM SPACEXTABLE WHERE Booster_Version = 'F9 v1.1';
```

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

AveragePayloadMass
2928.4

Explanation- The above picture shows result i.e. '2928.4' with query input for average Pay Load Mass when booster version was F9 v1.1.

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

```
]: %sql SELECT MIN(Date) AS FirstSuccessfulLandingDate FROM SPACEXTABLE WHERE Landing_Outcome = 'True RTLS';
```

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

]:

FirstSuccessfulLandingDate
None

Successful Drone Ship Landing with Payload between 4000 and 6000

Names of boosters which have successfully landed on drone ship and had payload mass greater tha
4000 but less than 6000

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 Booster_Version FROM SPACEXTABLE WHERE Landing_Outcome = 'True ASDS' AND PAYLOAD_MASS__KG_> 4000
```

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

Booster_Version

Task 7

Total Number of Successful and Failure Mission Outcomes

The total number of successful and failure mission outcomes

```
%sql SELECT SUM(CASE WHEN Landing_Outcome LIKE 'True%' THEN 1 ELSE 0 END) AS SuccessfulCount, SUM(CASE WHEN
```

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

SuccessfulCount	FailureCount
62	28

Explanation- Above are the given counts for success and failed launches of falcon 9.

Boosters Carried Maximum Payload

Names of the booster which have carried the maximum payload mass

```
In [33]: %sql SELECT Booster_Version FROM SPACEXTABLE WHERE PAYLOAD_MASS__KG_ = ( SELECT MAX(PAYLOAD_MASS__KG_) FROM
```

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

Out[33]:

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

- Here is the query for the 2015 Launch Records:

```
%%sql sqlite://
SELECT
     CASE
          WHEN substr(Date, 6, 2) = '01' THEN 'January'
          WHEN substr(Date, 6, 2) = '02' THEN 'February'
          WHEN substr(Date, 6, 2) = '03' THEN 'March'
          WHEN substr(Date, 6, 2) = '04' THEN 'April'
          WHEN substr(Date, 6, 2) = '05' THEN 'May'
          WHEN substr(Date, 6, 2) = '06' THEN 'June'
          WHEN substr(Date, 6, 2) = '07' THEN 'July'
          WHEN substr(Date, 6, 2) = '08' THEN 'August'
          WHEN substr(Date, 6, 2) = '09' THEN 'September'
          WHEN substr(Date, 6, 2) = '10' THEN 'October'
          WHEN substr(Date, 6, 2) = '11' THEN 'November'
          WHEN substr(Date, 6, 2) = '12' THEN 'December'
     END AS Month,
     Landing_Outcome,
     Booster_Version,
     Launch_Site
FROM SPACEXTABLE
WHERE substr(Date, 0, 5) = '2015'
AND Landing_Outcome LIKE 'False%'
AND Landing_Outcome LIKE '%ASDS%';
```

```
one.
```

Note- We used cases here for all the months in the year 2015

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

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

Out[37]:

Landing_Outcome	OutcomeCount
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

Expanation- We can observe that success and failure for drone ship is 5. Only one precluded that is for drone ship, for year 2015.

Section 3

Launch Sites Proximities Analysis

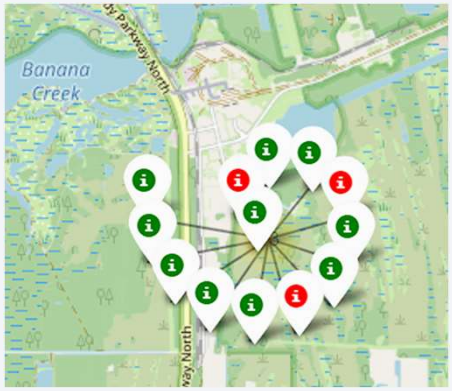
Where are the launch sites in the map?

Explore the generated folium map and make a proper screenshot to include all launch sites' location markers on a global map

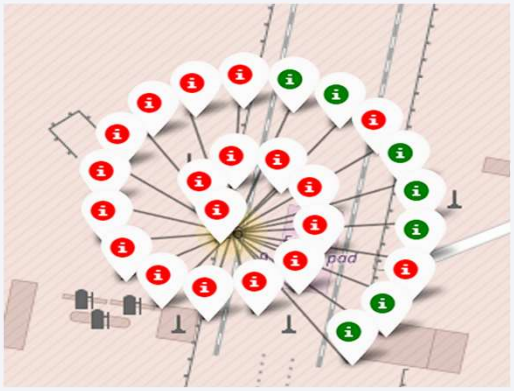
Explanation- First of all sites are in Florida in the right side and Los Angeles in left. We can see both the sites are chosen nearer the coastline area.

Launch Sites Near Florida

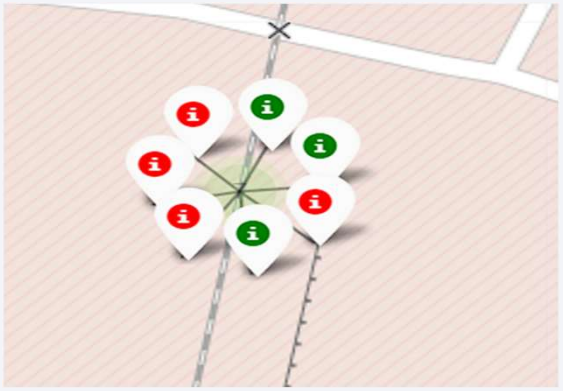
KSC LC-39A



CCAFS LC-40

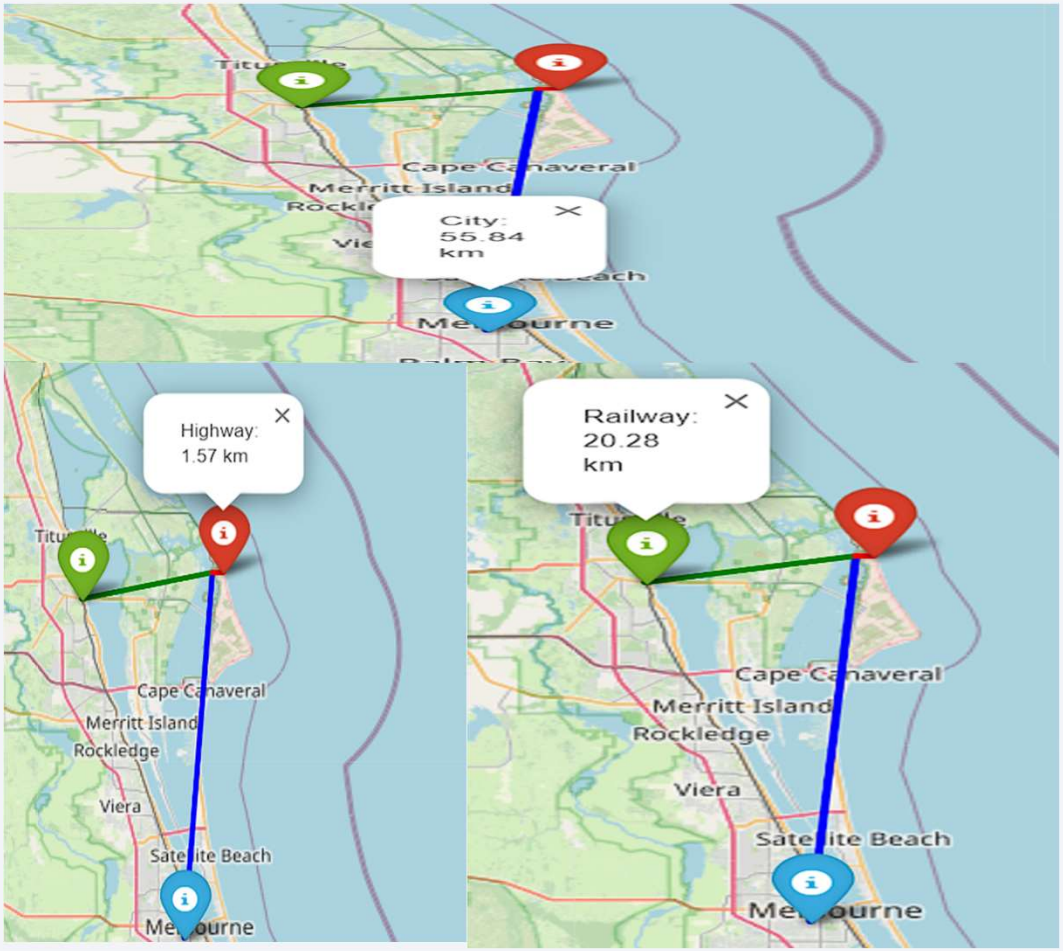


CCAFS SLC-40



Explanation- Green markers are the successful launches from the particular site and red for failed launch.

Representing near city, highway and railway from launch site:

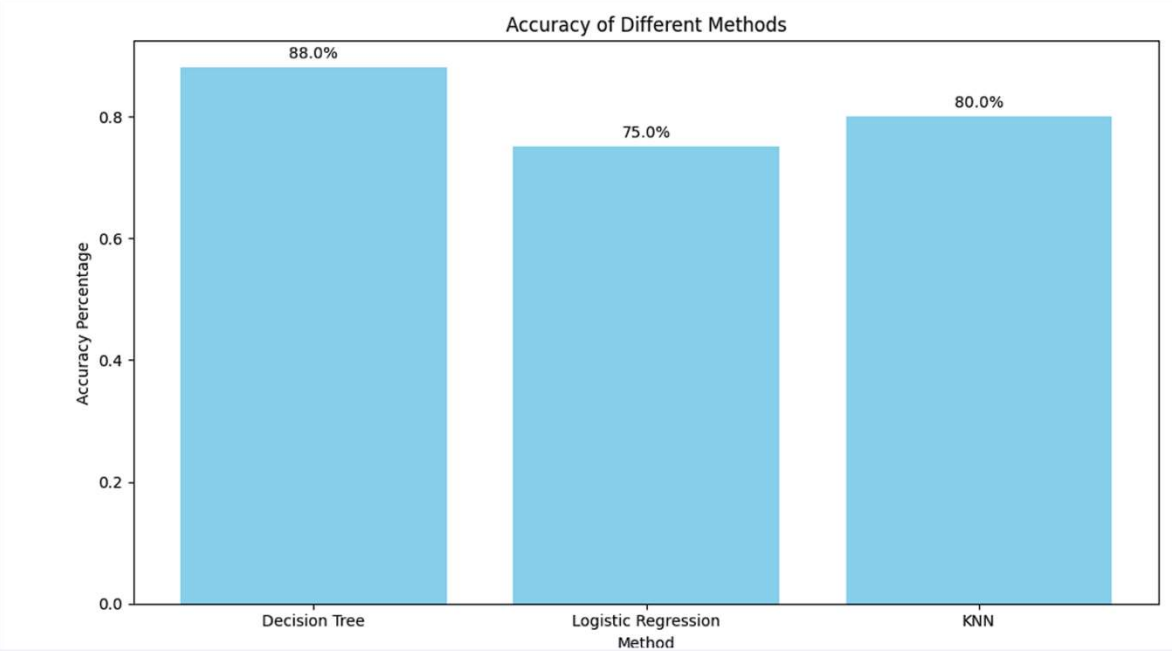


Section 5

Predictive Analysis (Classification)

Classification Accuracy

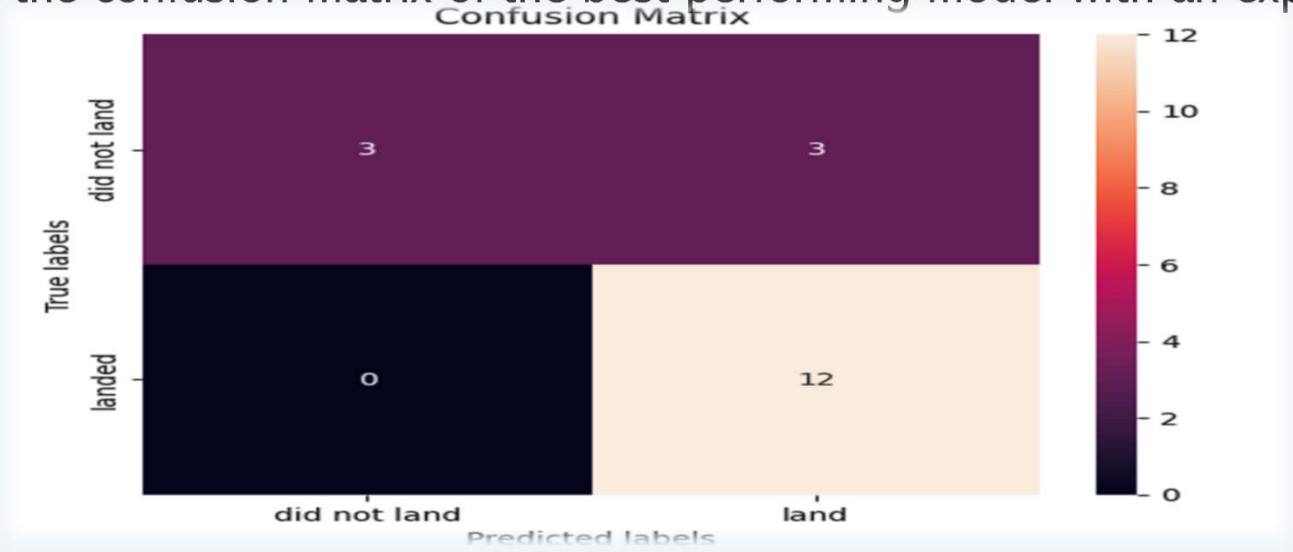
Visualize the built model accuracy for all built classification models, in a bar chart



Explanation- We have the best method which is Decision Tree having the accuracy 88% and least accuracy of 75%.

Confusion Matrix

Show the confusion matrix of the best performing model with an explanation



Description- **The ideal scenario is on the diagonal, where 12 landed labels were correctly predicted as landed and 2 did not land labels were correctly predicted as did not land. There were misclassifications though, with 3 landed labels being predicted as did not land and 10 did not land labels being predicted as landed. Overall, the model seems to be performing well because the majority of classifications (14) were correct, but it made some mistakes, particularly with misclassifying did not land labels.**

Conclusions

- Point 1- Classification Tree is the best predictive analysis for the acquired data of SPACEX Falcon9.
- Point 2- ESL1,GEO, HEO, SSO are the considerable orbits for future launchings.
- Point 3- Total Pay Load Mass Carried by boosters from NASA is 45596 KG.
- Point 4- Average Pay Load Mass from booster version F9 v1.1 is 2928.4 KG.
- Point 5- When Launch Site was CCAFS LC-40 orbit always got chosen as LEO resulted success.
- Point 6- Although Classification Tree method is best in all but there is always a chance to maximize it.

Appendix

- We have faced issue in the marking cluster part and also the representation part of distance calculated on map. We decoded the whole jupyter file again and then got to know where we did mistake.
- We used KNN, Logistic Regression, SVM and Classification Tree approach in the predictive analysis.
- Best Method was Classification Tree for the provided SPACEX Data.
- I used word ‘We’ several times in the report but it actually means to ‘I’ but not used because it does not feel like professional.

Thank you!