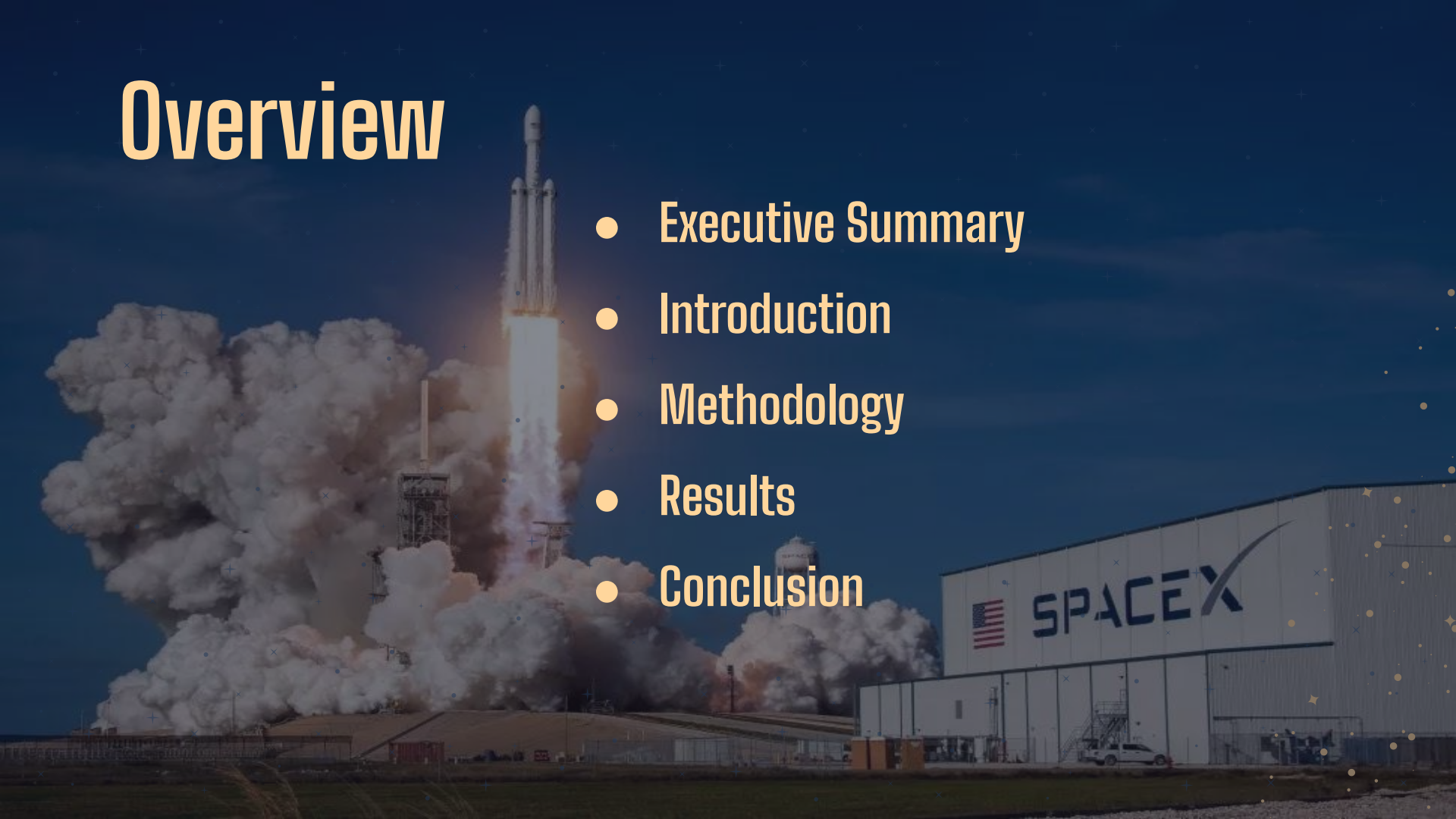
SpaceX Falcon 9 First Stage Landing Prediction Project

Jinjing Zhang 09.18.2021
(jinjingzhang61@gmail.com)

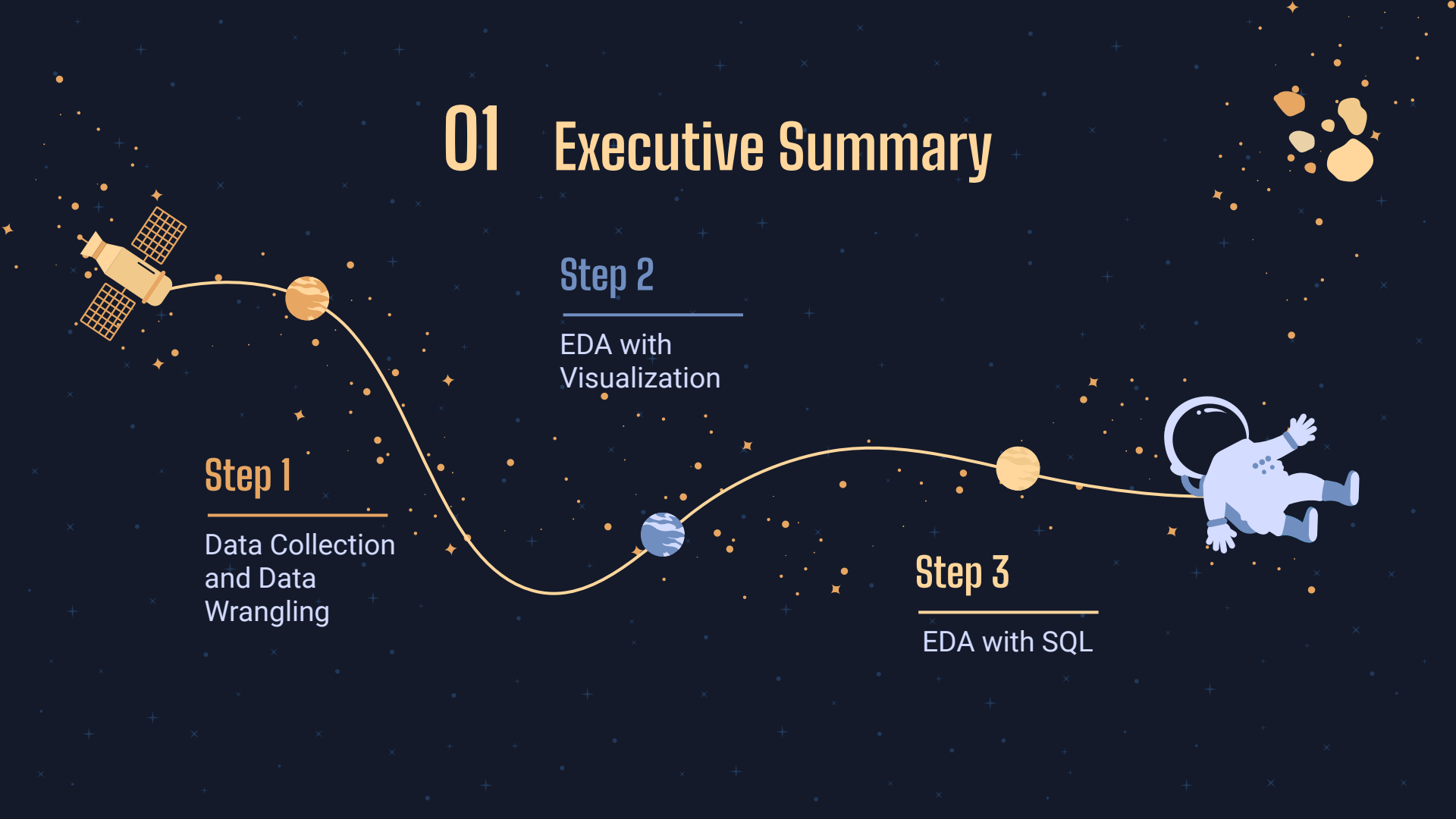
Guided by IBM Corporation

Overview



- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

01 Executive Summary



Step 1

Data Collection and Data Wrangling

Step 2

EDA with Visualization

Step 3

EDA with SQL

Step 4

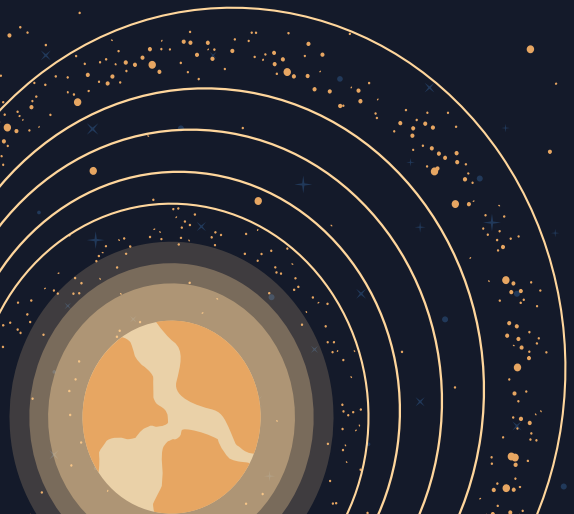
Interactive Map with Folium

Step 5

Dashboard with Ploty Dash

Step 6

Predictive Analysis (Classification)



02

INTRODUCTION

02

INTRODUCTION

In this project, we will predict whether the Falcon 9 first stage will successfully land. SpaceX advertises on its website that the Falcon 9 rocket cost $62 million to launch; other suppliers cost up to $165 million, with most of the savings due to SpaceX's ability to reuse the first stage. So if we can determine whether the first stage will land, we can determine the cost of the launch. This information could be used if another company wanted to bid on a rocket launch with SpaceX.

03

Methodology

3.1 API Data Collection


```
url =”https://api.spacexdata.com/v4/capsules”
url =”https://api.spacexdata.com/v4/cores”
url =”https://api.spacexdata.com/v4/launches/past”

Response = requests.get ( url )
Response.json ( )
df = pd.json_normalize( Response )
```

API is the acronym for Application Programming Interface, which is a software intermediary that allows two applications to talk to each other.

These URLs will be used to target a specific endpoint of the API to get past launch data. Then we decode the response content as a Json using .json() and turn it into a Pandas dataframe using .json_normalize()

3.2 API Data Wrangling


```
def getLaunchSite(data):
    for x in data['launchpad']:
        response =
requests.get("https://api.spacexdata.com/v4/launchpads/"+str(x)).jso
n()
        Longitude.append(response['longitude'])
        Latitude.append(response['latitude'])
        LaunchSite.append(response['name'])
```

We create several similar functions to collect a total of 17 columns, including 'FlightNumber', 'Date', 'BoosterVersion', 'PayloadMass', 'Orbit', "LaunchSite', 'Outcome', 'Longitude', 'Latitude', etc.

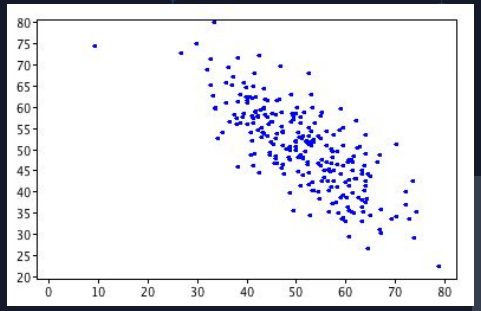
3.2 API Data Wrangling

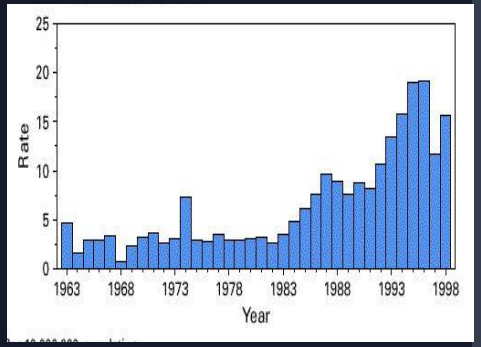
03

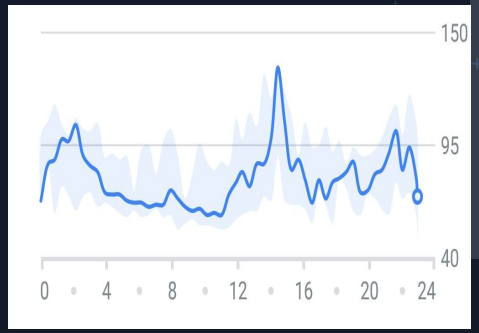
Methodology

The data should be sufficiently cleaned before data analysis is performed. We filter out unnecessary rows, remove 'null value', or replace 'null value' with mean value.

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude	Class
1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003	-80.577366	28.561857	0
2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005	-80.577366	28.561857	0
3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007	-80.577366	28.561857	0
4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003	-120.610829	34.632093	0
5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004	-80.577366	28.561857	0







3.3 Data Visualization

1. Scatter Plot: is used to observe relationships between variables.
    i. Payload vs Flight Number
    ii. LaunchSite vs Flight Number
    iii. LaunchSite vs Payload
    iv. Orbit Type vs Flight Number
    v. Payload vs Orbit Type
2. Bar Plot: is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent.
    Launch Success Rate vs Orbit Type
3. Line Plot: is a way to display data along a number line.
    Launch Success Rate Yearly Trend

3.4 EDA With SQL

SQL (Structured Query Language) is a standardized programming language that's used to manage relational databases and perform various operations on the data in them.

In this study, we will use SQL in 'Jupyter Notebook' to query data from given tables stored in 'IBM Cloud DB2' and use SQL to clean the data.

```
!pip install sqlalchemy==1.3.9
!pip install ibm_db_sa
!pip install ipython-sql

%load_ext sql
```

```
%sql ibm_db_sa://cln08992:8hplOWVtXQtO2JJI
```

```
%%sql
select * from SPACEXTBL
limit 20
```

The image on the left is a sample SQL code from the Juypter Notebook connected to the IBM online database.
The first 20 rows of TABLE 'SPACEXTBL' are selected.

3.4 EDA With SQL

What do we need to query with SQL in this study?

1. Display the names of the unique launch sites in the space mission.
2. Display 5 records where launch sites begin with the string 'CCA'.
3. Display the total payload mass carried by boosters launched by NASA (CRS).
4. Display average payload mass carried by booster version F9 v1.1.
5. List the date when the first successful landing outcome in ground pad was acheived.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
7. List the total number of successful and failure mission outcomes.
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery.
9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.
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.

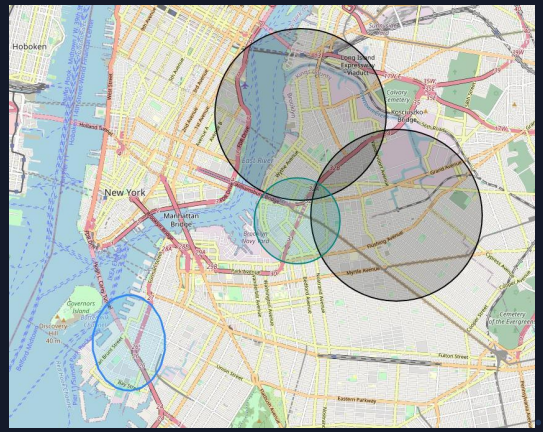
3.5 Build An Interactive Map With Folium

(Geospatial Analysis)

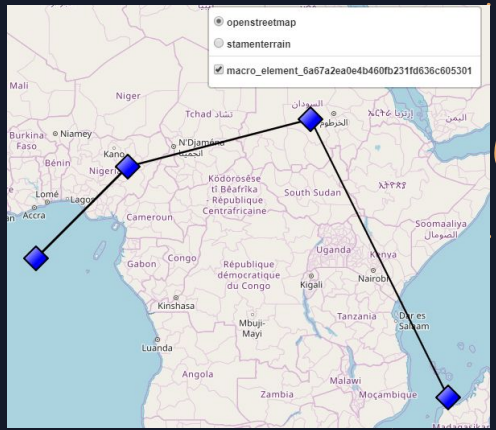
The Folium library in Python helps us analyze the location and geospatial data with ease, and lets us create interactive maps.

Folium Map Marker

Folium Circle Marker



Folium PolyLine



Resource: Folium Map Marker https://media.geeksforgeeks.org/wp-content/uplds/volcanoes.jpg; Folium Circle Marker https://i.stack.imgur.com/ATIR6.jpg; Polyline: https://i.stack.imgur.com/jG0e4.png

3.6 Interactive Dashboard With Plotly Dash

Dash is a python framework created by plotly for creating interactive web applications.

Callbacks in Dash: A callback is initialised using @app.callback(), which is followed by a function definition. Within this function, we define what happens on changing the value of the dropdown.

Some sample code:

```
import dash
import dash_html_components as html
import dash_core_components as dcc
from dash.dependencies import Input, Output
import pandas as pd
import plotly.express as px
```

```
dcc. Dropdown ( ) # Create a dropdown
dcc.RangeSlide ( ) # Create a rangeslide
px.pie ( ) # Create a pie plot
px.scatter ( ) # Create a scatter plot
```

3.7 Classification (Predictive Analysis)

The following are four machine learning classification models

Logistic Regression

Logistic regression is a calculation used to predict a binary outcome: either something happens, or does not.

Decision Tree

A decision tree is a supervised learning algorithm that is perfect for classification problems, as it's able to order classes on a precise level.

SVM

A support vector machine (SVM) uses algorithms to train and classify data within degrees of polarity, taking it to a degree beyond X/Y prediction.

KNN

K-nearest neighbors (k-NN) is a pattern recognition algorithm that uses training datasets to find the k closest relatives in future examples.

3.7 Classification (Predictive Analysis)

The following are some methods will be used in the predictive analysis.

Train-Test Split

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

GridSearchCV

GridSearchCV is a library function that is a member of sklearn's model_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.

Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem.

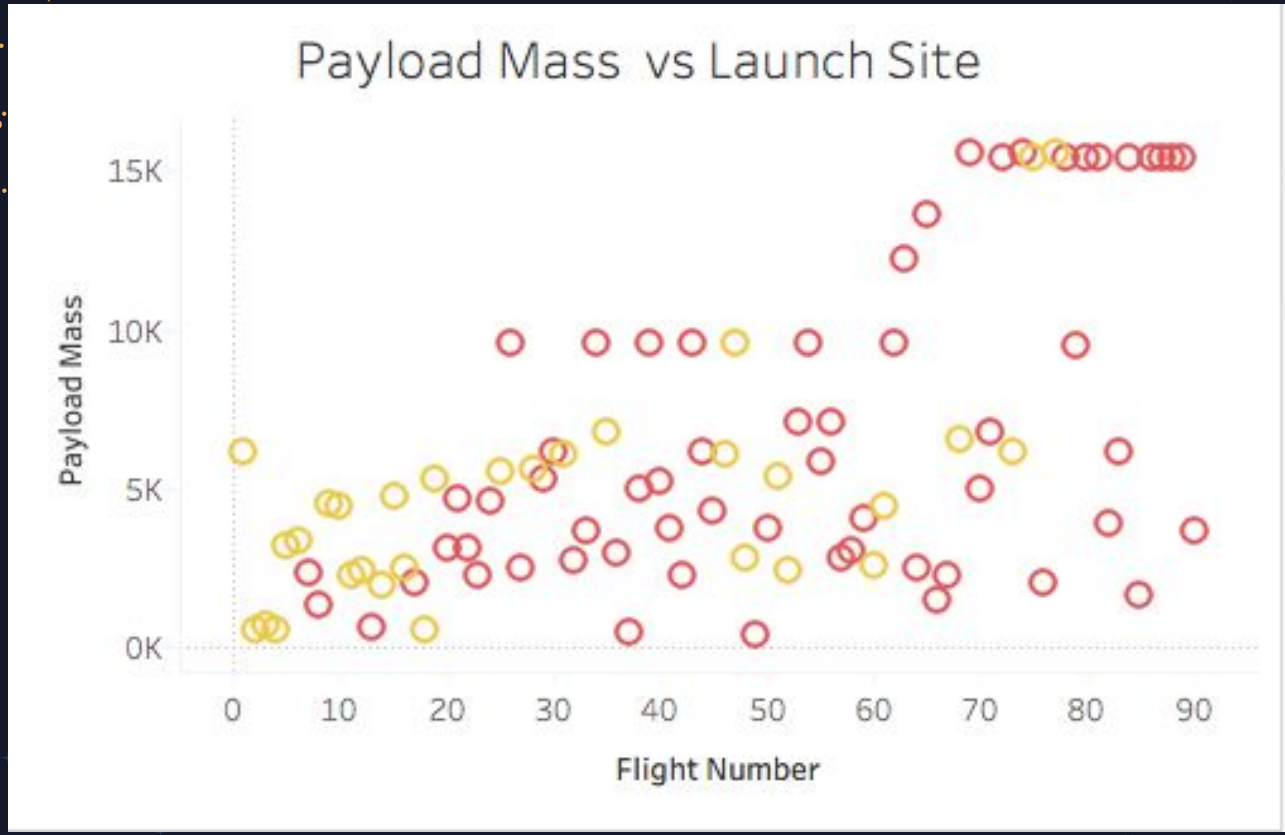
It presents results including True positive (TP), True negative (TN), False positive (FP), False positive (FP).

Resource: Train-Test Split, ttps://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/ ; Grid Search for Hyperparameter Tuning, https://towardsdatascience.com/grid-search-for-hyperparameter-tuning-9f63945e8fec

04

Results

4.1 EDA With Visualization Results



Launch Success Rate

0

1

Summary

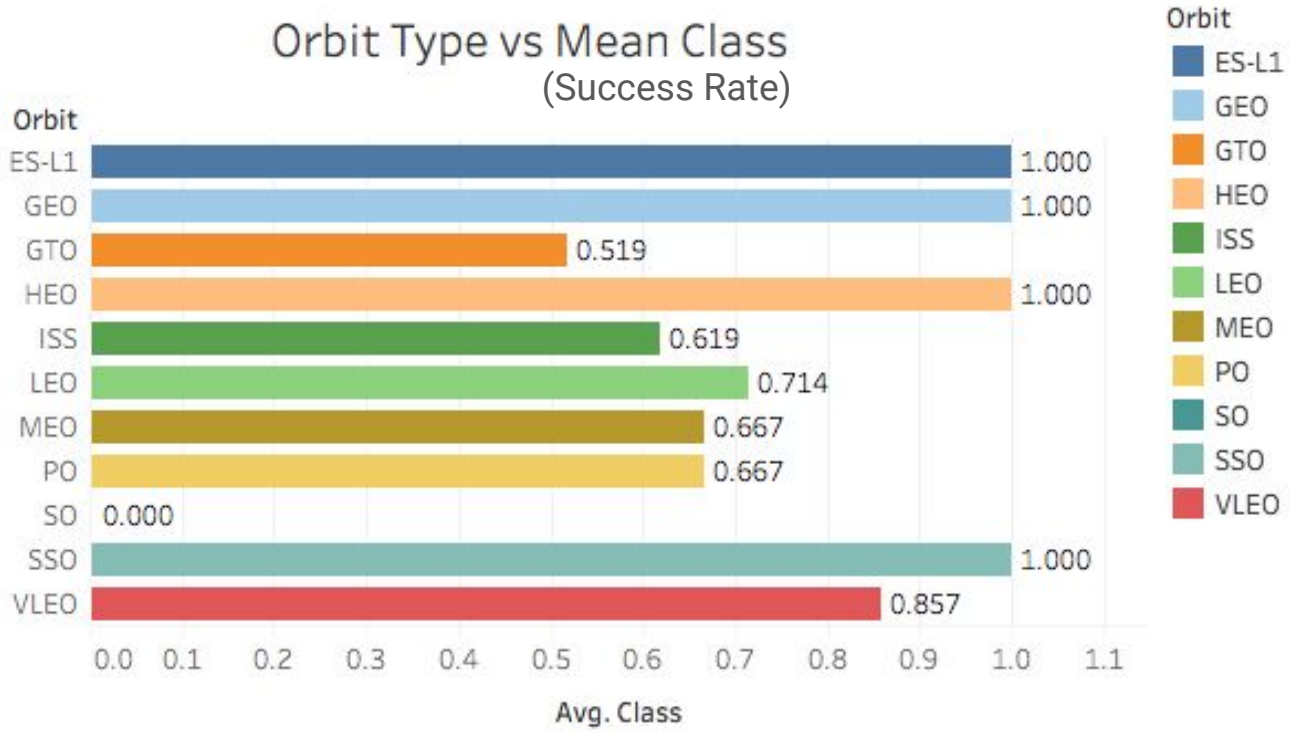
Count: 90

We see that as the flight number increases, the first stage is more likely to land successfully. The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

4.1 EDA With Visualization Results

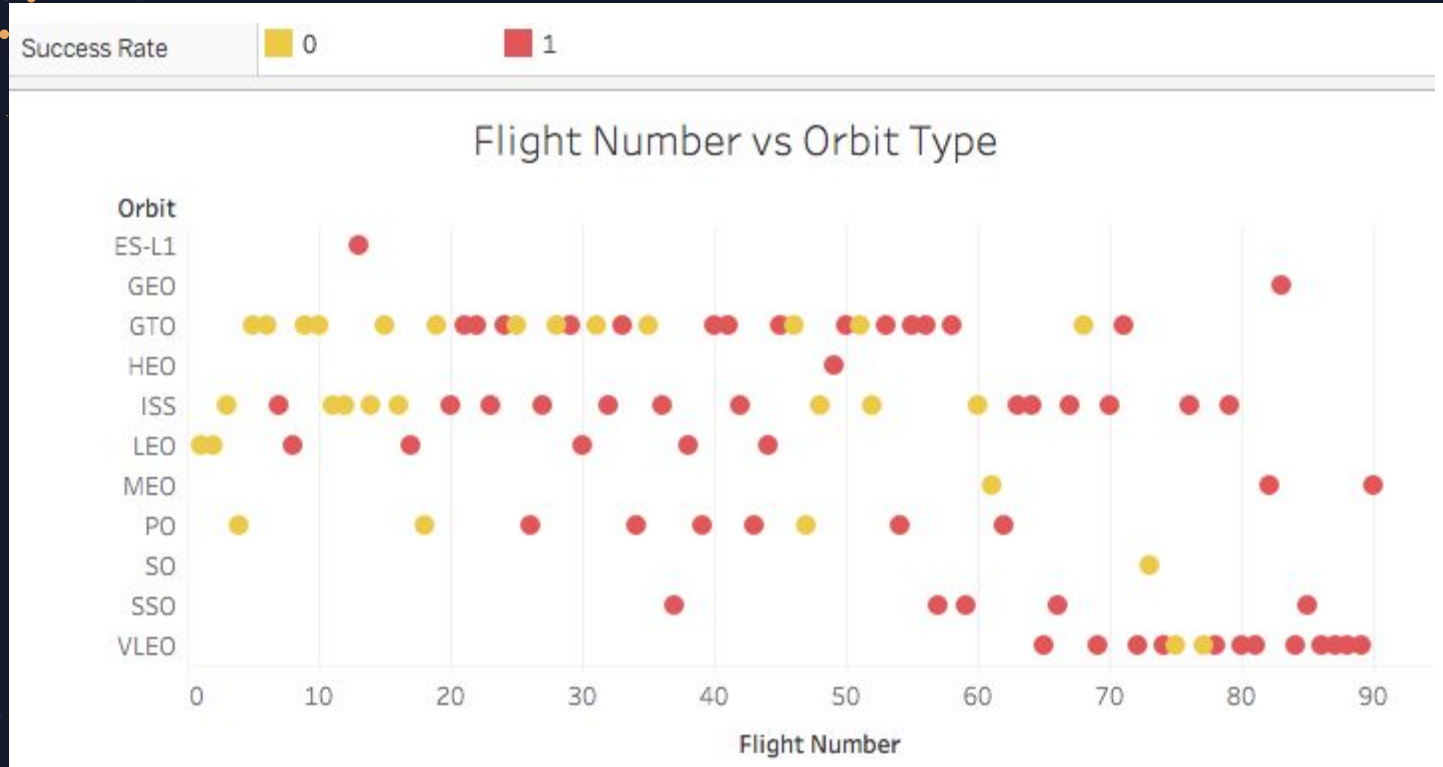
The Flight Number increases from small to large, corresponding to the rocket launch time from early to late. It can be seen from the graph that the first stage launch success rate increases as the Flight Number increases. In addition, the highest number of flights came from the launch site of 'CCAFS SLC 40'.

4.1 EDA With Visualization Results



We can find that ES-L1, GEO, GTO have highest launch success rate 100%, and SO has lowest success rate 0%.

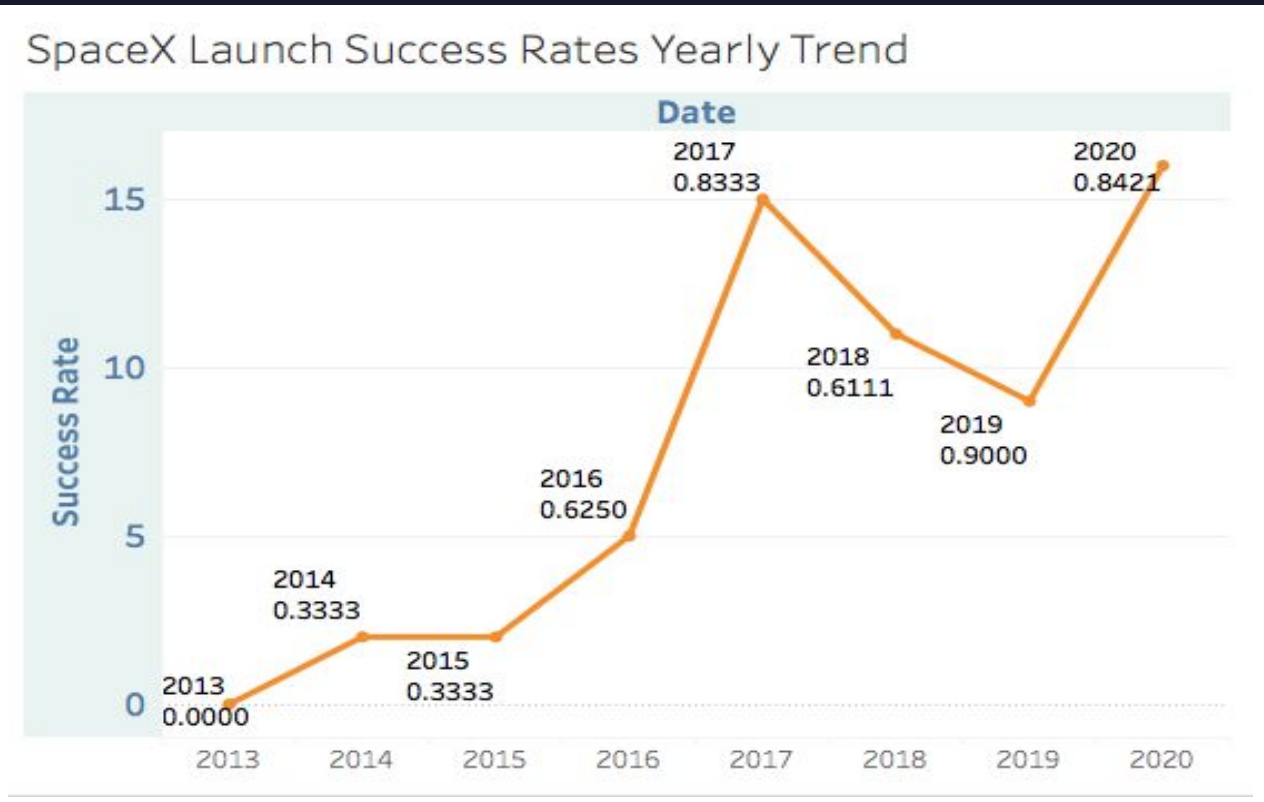
4.1 EDA With Visualization Results



Each launch is aim towards an dedicated orbit, most of the recently launched rockets (with a large Flight Number) are aim to 'VLEO' orbit and have a high success rate.

4.1 EDA With Visualization Results

We can observe that the success rate of rocket first stage launch continues to rise from 2013 to 2020, and although it tends to fall after 2017, it reaches a peak in 2020.



4.2 EDA With SQL results

Q1: Display the names of the unique launch sites in the space mission.

SQL Code:

```
%sql SELECT DISTINCT launch_site from SPACEXTBL
```

Query Results:

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

4.2 EDA With SQL results

Q2: Display 5 records where launch sites begin with the string 'CCA'.

SQL Code:

```
%%sql
select * from SPACEXTBL
   where launch_site like '%CCA%'
   limit 5
```

Query Results:

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	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	00:35:00	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

4.2 EDA With SQL results

Q3: Display the total payload mass carried by boosters launched by NASA (CRS).

SQL Code:

```
%%sql
SELECT SUM(payload_mass__kg_) total_NASA_CRS FROM SPACEXTBL
   WHERE customer = 'NASA (CRS)'
   GROUP BY customer
```

Query Results:

total_nasa_crs
45596

4.2 EDA With SQL results

Q4: Display average payload mass carried by booster version F9 v1.1.

SQL Code:

```
%%sql
SELECT AVG(payload_mass__kg_) AVG_F9, booster_version
        FROM SPACEXTBL
        WHERE booster_version LIKE '%F9 v1.1%'
        GROUP BY booster_version
```

Query Results:

avg_f9	booster_version
2928	F9 v1.1
500	F9 v1.1 B1003
2216	F9 v1.1 B1010
4428	F9 v1.1 B1011
2395	F9 v1.1 B1012
570	F9 v1.1 B1013
4159	F9 v1.1 B1014
1898	F9 v1.1 B1015
4707	F9 v1.1 B1016
553	F9 v1.1 B1017
1952	F9 v1.1 B1018

4.2 EDA With SQL results

Q5: List the date when the first successful landing outcome in ground pad was acheived.

SQL Code:

```
%%sql
SELECT MIN(Date) AS first_landing, landing__outcome FROM SPACEXTBL
   WHERE landing__outcome LIKE '%Success (ground pad)%'
   GROUP BY landing__outcome
```

Query Results:

first_landing	landing__outcome
2015-12-22	Success (ground pad)

4.2 EDA With SQL results

Q6: List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

SQL Code:

```
%%sql
SELECT booster_version, landing__outcome,payload_mass__kg_ FROM SPACEXTBL
    WHERE landing__outcome LIKE '%Success (drone ship)%'
    AND payload_mass__kg_ BETWEEN 4000 AND 6000
```

Query Results:

booster_version	landing__outcome	payload_mass__kg_
F9 FT B1022	Success (drone ship)	4696
F9 FT B1026	Success (drone ship)	4600
F9 FT B1021.2	Success (drone ship)	5300
F9 FT B1031.2	Success (drone ship)	5200

4.2 EDA With SQL results

Q7: List the total number of successful and failure mission outcomes.

SQL Code:

```
%%sql
SELECT COUNT(mission_outcome) counts, mission_outcome FROM SPACEXTBL
        GROUP BY  mission_outcome
```

Query Results:

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

4.2 EDA With SQL results

Q8: List the names of the booster_versions which have carried the maximum payload mass. Use a subquery.

SQL Code:

```
%%sql
SELECT booster_version, payload_mass__kg_
        FROM SPACEXTBL
        WHERE payload_mass__kg_ =
                (SELECT  MAX ( payload_mass__kg_ )
                FROM SPACEXTBL)
```

Query Results:

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

4.2 EDA With SQL results

Q9: List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015.

SQL Code:

```
%%sql
SELECT DATE, booster_version, launch_site, landing__outcome
    FROM SPACEXTBL
    WHERE
    DATE BETWEEN '2015-01-01' AND '2015-12-31'
```

Query Results:

DATE	booster_version	launch_site	landing__outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-02-11	F9 v1.1 B1013	CCAFS LC-40	Controlled (ocean)
2015-03-02	F9 v1.1 B1014	CCAFS LC-40	No attempt
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)
2015-04-27	F9 v1.1 B1016	CCAFS LC-40	No attempt
2015-06-28	F9 v1.1 B1018	CCAFS LC-40	Precluded (drone ship)
2015-12-22	F9 FT B1019	CCAFS LC-40	Success (ground pad)

4.2 EDA With SQL results

Q10: 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.

SQL Code:

```
%%sql
SELECT COUNT ( * ) counts, landing__outcome FROM
(SELECT * FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20')
GROUP BY landing__outcome
ORDER BY counts DESC
```

Query Results:

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

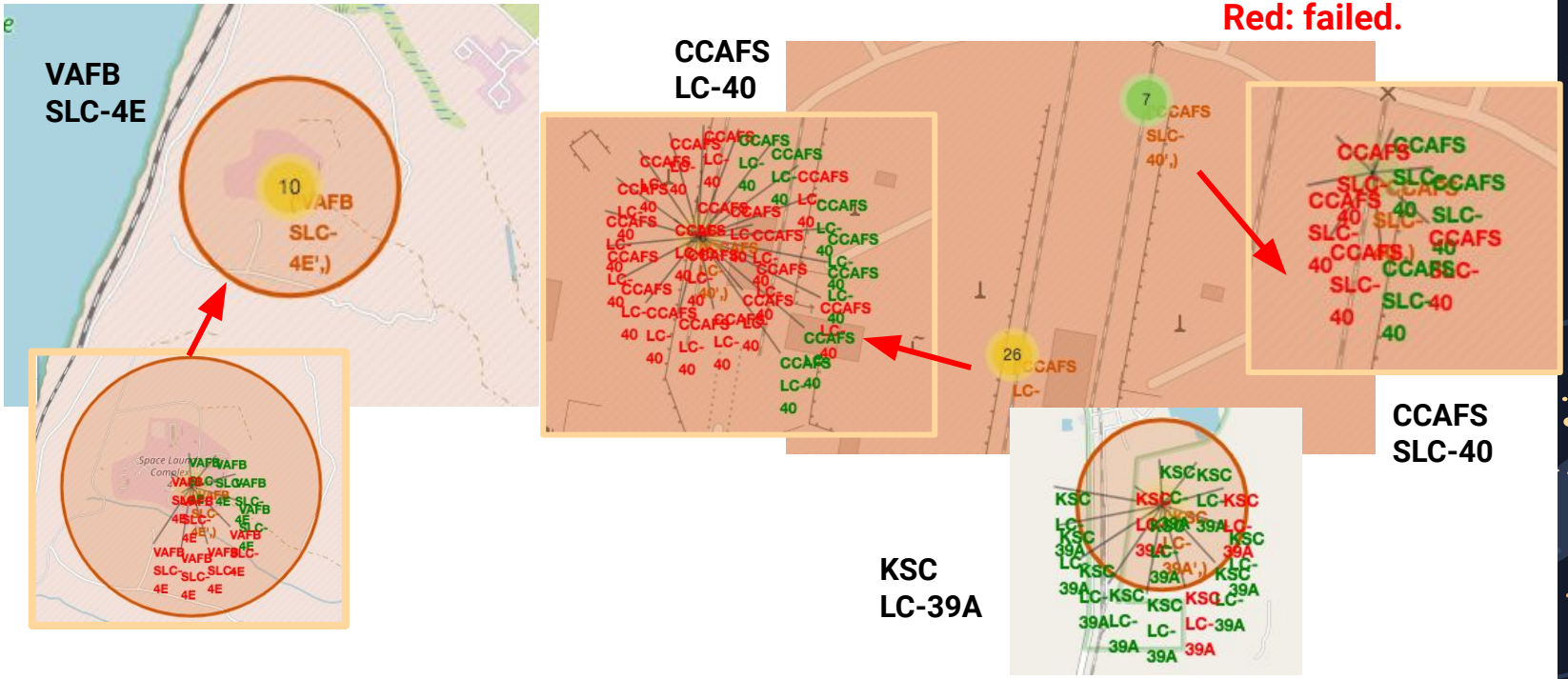
4.3 Interactive Map With Folium Results

Launch SItes on Folium Map

4.3 Interactive Map With Folium Results

Mark the success/failed launches for each site on the map.

Green: success,
Red: failed.



4.3 Interactive Map With Folium Results

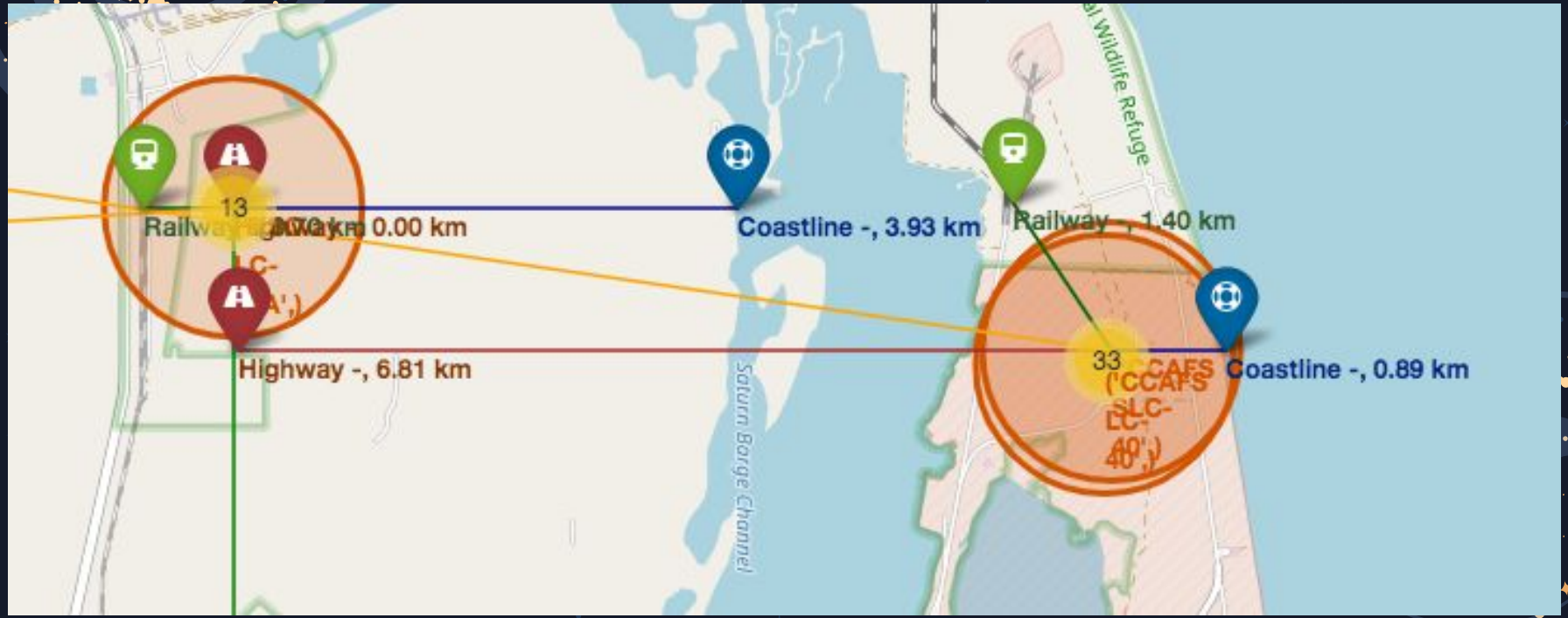
Distance From Launch Sites to Equator

4.3 Interactive Map With Folium Results

Distance from launch sites to subway, highway, coastline

4.3 Interactive Map With Folium Results

Distance from launch sites to subway, highway, coastline



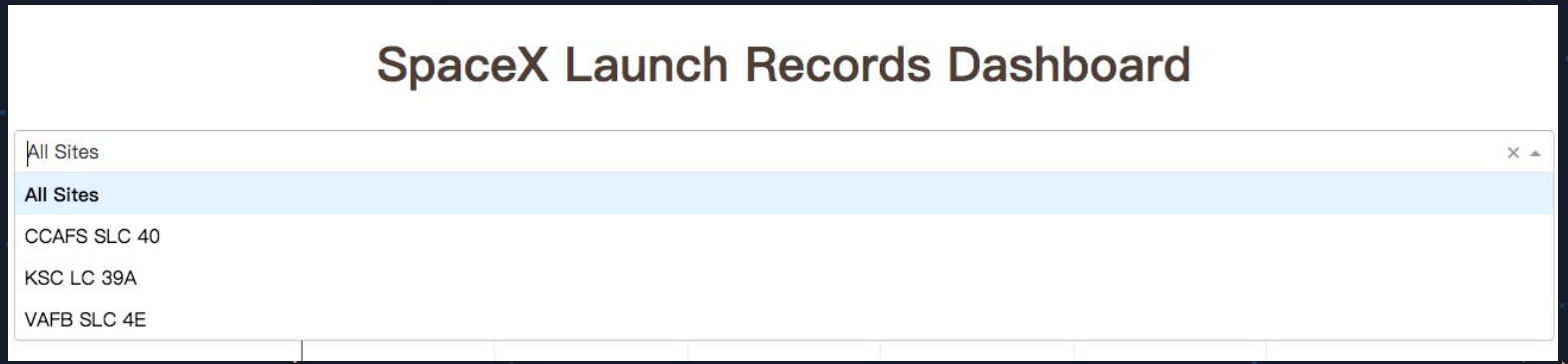
4.3 Interactive Map With Folium Results

Conclusion

1. All launch sites are far from the equator line. (> 3000 km)
2. All launch sites are close to railways. (< 2km)
3. All launch sites are not far from haiways. (< 15km)
4. All launch sites are close to coastline. (< 5km)
5. All launch sites are keep away from cities. (> 15km)

4.4 Interactive Plotly Dash Dashboard Results

A drop-down menu is used to filter the Launch Sites.



jinjingzhang-8050.theiadocker-2-labs-prod-theiak8s-3-tor01.proxy.cognitiveclass.ai

应用 jinjingzhang61@g... jinjingzhang61 Go... 阅读清单

SpaceX Launch Records Dashboard

All Sites

Total Success Launches by All Sites

KSC LC-39A Launch Site has the highest success rate: 41.7%.

Payload range (Kg):

Correlation Between Payload and Success for All Sites

jinjingzhang-8050.theiadocker-2-labs-prod-theiak8s-3-tor01.proxy.cognitiveclass.ai

应用 jinjingzhang61@g... jinjingzhang61 Go... 阅读清单

SpaceX Launch Records Dashboard

VAFB SLC-4E

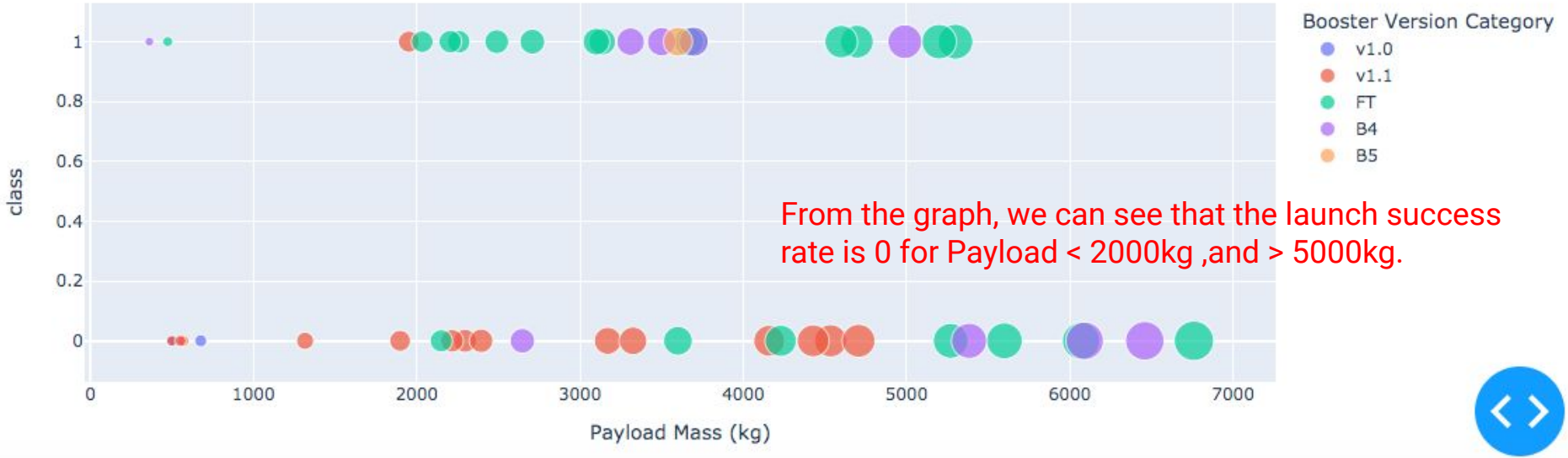
Total Success Launches for Site → VAFB SLC-4E

Payload range (Kg):

Correlation Between Payload and Success for Site → VAFB SLC-4E

Payload range (Kg):

Correlation Between Payload and Success for All Sites



From the graph, we can see that the launch success rate is 0 for Payload < 2000kg ,and > 5000kg.

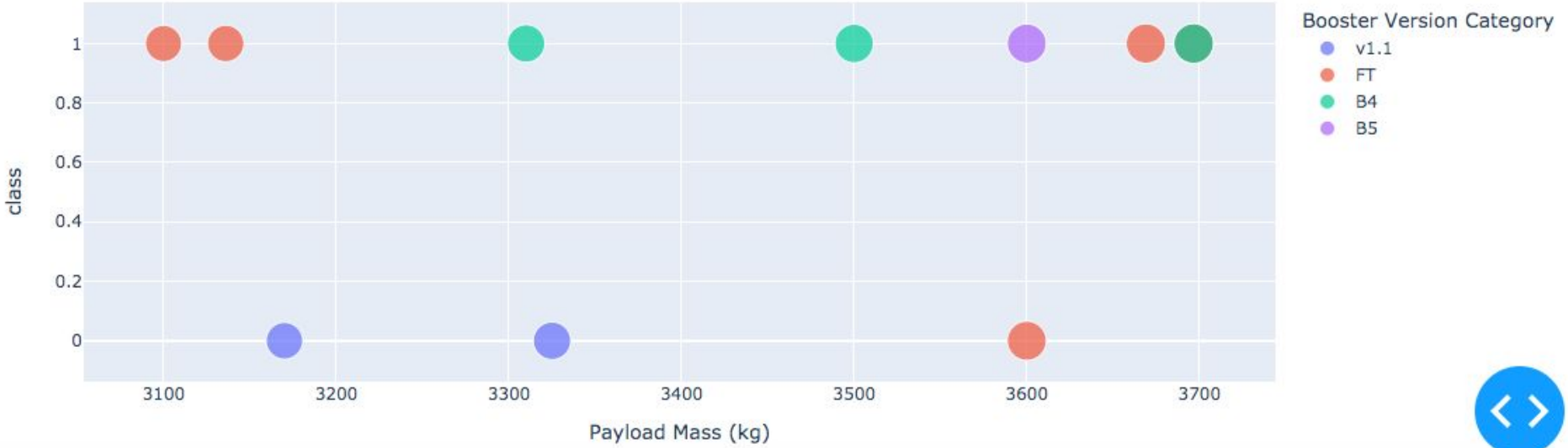
Payload range (Kg):

Payload in the range of 6000 kg - 7000 kg have a 0 success rate.

Payload range (Kg):

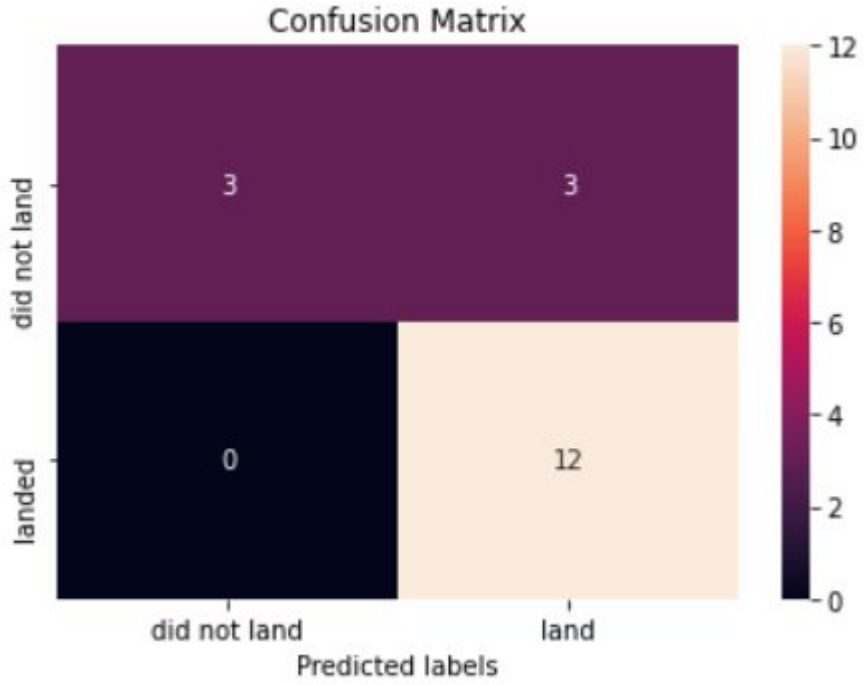
Payload in the range of 3000 kg - 4000 kg have a relatively high success rate.

Correlation Between Payload and Success for All Sites



Select to present any one or more Booster Versions only

4.5 Predictive Analysis (Classification) Results

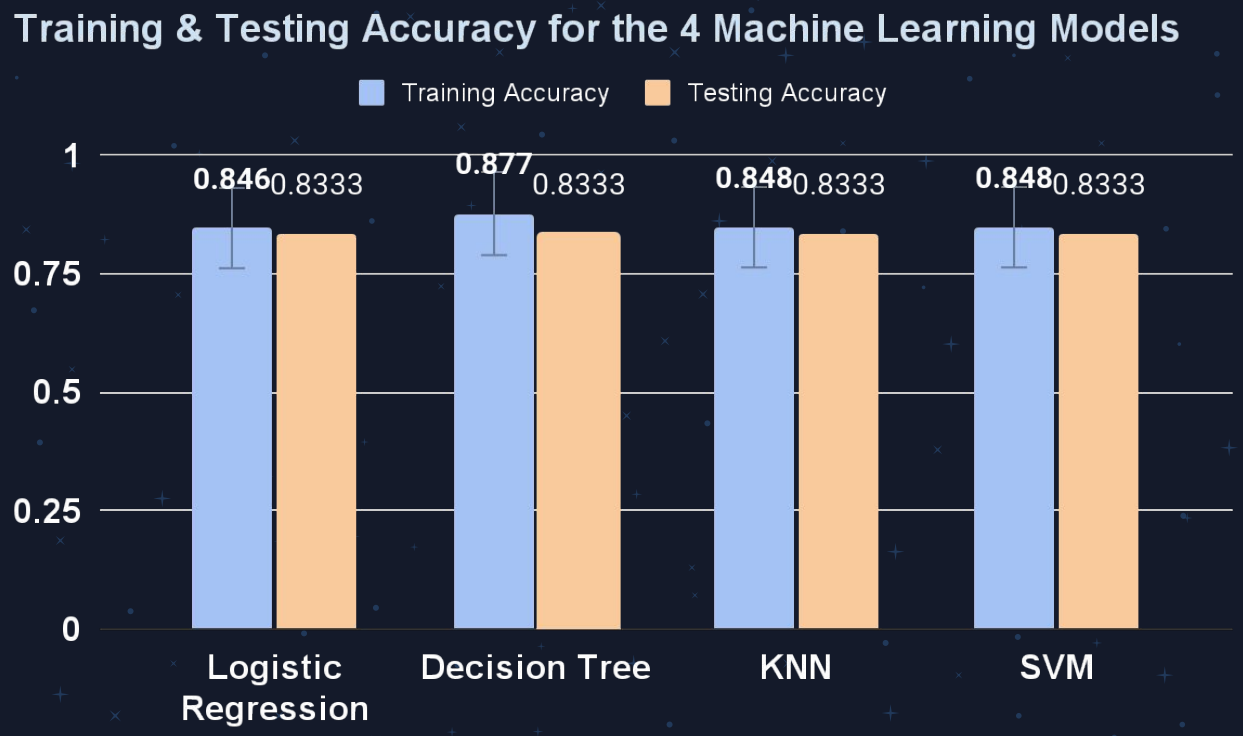


The four models obtain the same Confusion Matrix
True Positive (TP) =12
False Positive (FP) = 3
True Negative (TN) = 3
False Negative (FN) = 0

True Positive Rate: TP/ (TP + FN) = 1
False Positive Rate: FP/ (FP + TN) = 0.5
True Negative Rate: TN/ (TN + FP) = 0.5
False Negative Rate: FN/ (TP + FN) = 0

Accuracy: (TP + TN)/TOTAL = 0.833

4.5 Predictive Analysis (Classification) Results



The method which performs best is "Decision Tree" with a traning accuracy of 0.8767857142857143

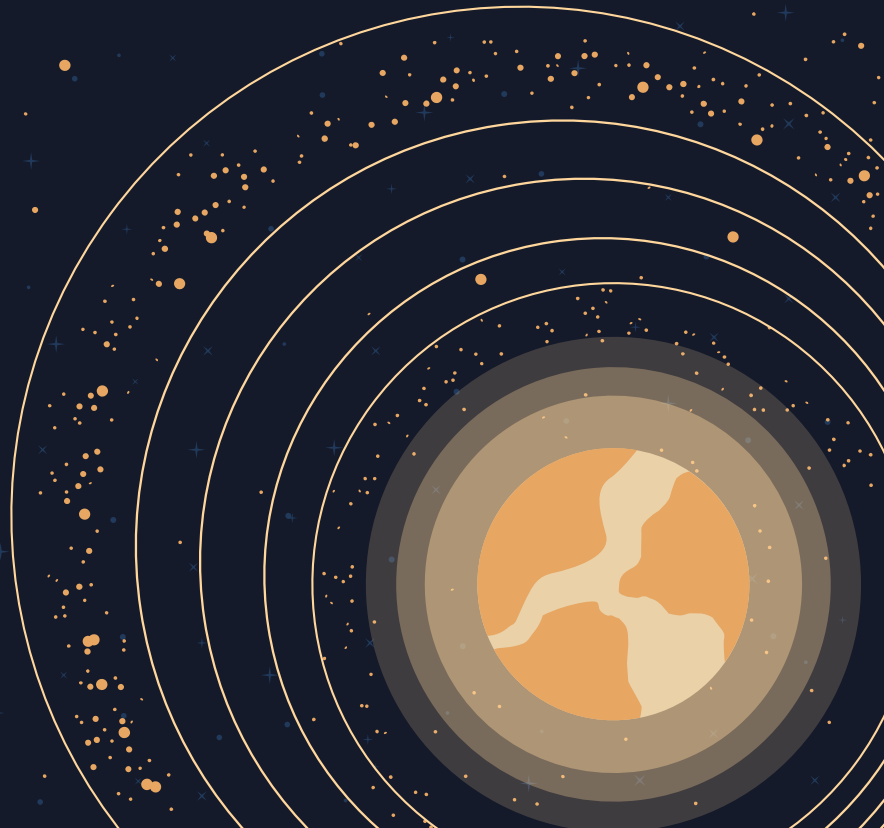
4.5 Predictive Analysis (Classification) Results

Tuned hyperparameters (best parameters)

- **Logistic Regression:** {'C': 0.01, 'penalty': 'l2', 'solver': 'lbfgs'}
- **SVM:** {'C': 1.0, 'gamma': 0.031622776601683793, 'kernel': 'sigmoid'}
- **Decision Tree:** {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'best'}'
- **KNN**: {'algorithm': 'auto', 'n_neighbors': 10, 'p': 1}

05

Conclusion



Conclusion

1. The success rate of rocket(first stage) launch continues to rise from 2013 to 2020, and although it tends to fall after 2017, it reaches a peak in 2020.
2. The majority of recent rocket launches have aimed towards the "VLEO" orbit with a success rate of 85.7%.
3. Orbit ‘ES-L1’, ‘GEO’, ‘GTO’ have highest launch success rate 100%, and ‘SO’ has lowest success rate 0%.
4. ‘KSC LC-39A’ launch site has the highest launch success rate.
5. Launch sites are close to coastline and railways.
6. As the payload mass becomes very larger, the launch success rate decreases.
7. Payload in the range of 3000 kg - 4000 kg have a relatively high success rate.
8. The best performing method in predictive analysis is the "Decision Tree" with a training accuracy of 0.877 and a testing accuracy of 0.833.

THANK YOU!