IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Alalade Oluwaferanmi, O

24th August, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data about Falcon 9 launches was acquired from public sources – SpaceX data site and Wikipedia;
  - Research was performed primarily with Python libraries for data acquisition, wrangling, visualization and modelling
  - Correlations between various parameters (time, site, orbit, payload) and landing success were studied, 4 models were built and tested;
- Summary of all results
  - Models built show predictive accuracy for landings outcome between 83 and 95 percent, which rather high. The highest score produced Decision Tree model; •
  - There are major correlations between:
  - time (failures become rarer),
  - orbit (launches to low orbits have higher landing success rate),
  - payload (high-loaded launches are more successful in landings, though it may derive from orbits correlation)

Introduction

- The Falcon 9 is a vehicle that can carry cargo and humans into Space. The rocket has two stages.
  - The Booster Stage: Which carries the second stage to a certain altitude, which then returns to the earth surface, either on Drone ships, Ground pads, to facilitate reuse.
  - The second Stage: which carries the cargo to the destination.
- This project is to predict the Falcon 9 landing, if the first stage will land successfully or not using Machine Learning algorithms.

Section 1

Methodology

Methodology

Executive Summary

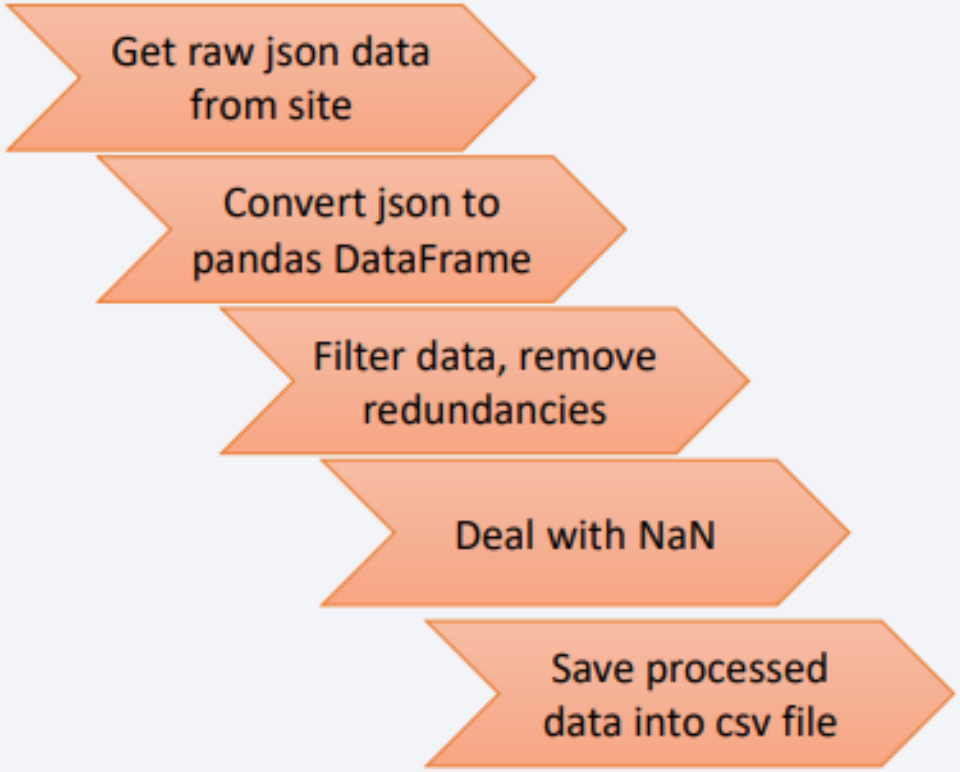
- Data collection methodology:
  - SpaceX website: https://docs.spacexdata.com/
  - Wikipedia: List of Falcon 9 and Falcon Heavy launches - Wikipedia
- Perform data wrangling
  - Replacement of missing values
  - Representation of landing outcomes as binary
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Using Grid Search to find the best parameters for each model

Data Collection

- The data used in this project was gotten from the SpaceX website itself and Wikipedia. Additional information was gotten from Wikipedia also for the description and full forms of abbreviations

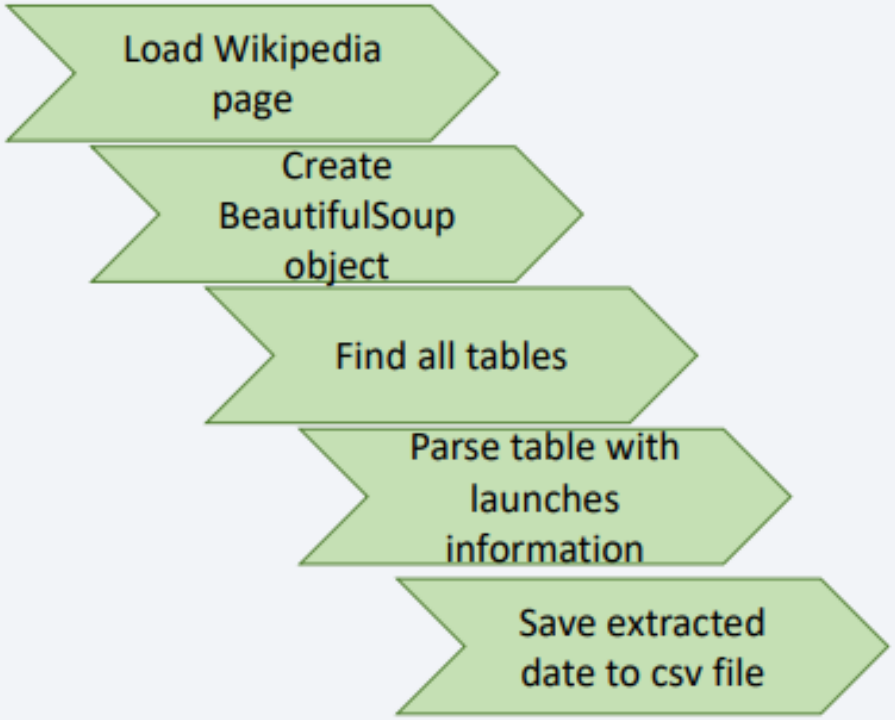
Data Collection – SpaceX API

- The data was collected using REST APIs from the spacexdata API using requests.
- https://docs.spacexdata.com/
- It was then turned to a DataFrame using json_normalize. The resulting DataFrame consisted of 43 columns which was then narrowed down into six which were Rocket, Payloads, Launchpad, Cores, flightnumber, date_utc. These columns were then used to get additional information using predefined functions. The final DataFrame called *df* consisted of 17 columns namely FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude and Latitude.
- The DataFrame *df* was then filtered to contain only the 'Falcon 9' rockets in the BoosterVersion column.
    - https://github.com/Feranmi-Alalade/IBM-Data-Science-Capstone-SpaceX/blob/main/1_Data_Collection.ipynb



Data Collection - Scraping

- The Wikipedia was first loaded and the request method was used.
- A beautiful soup object was created and using the find_all() method, all tables were found using the keyword “Table”, the information in each table was then parsed into a dictionary by using the same find_all() method using “th” and “tr”.
- The data was then extracted into a csv file



- https://github.com/Feranmi-Alalade/IBM-Data-Science-Capstone-SpaceX/blob/main/2_Webscraping.ipynb

Data Wrangling

- Acquired data was processed with Pandas library
- To determine initial patterns following values were calculated:
  - number of launches by site;
  - number of occurrence of each orbit;
  - number of different landing outcomes;
- To enable further analysis landing outcomes were categorized into 0 (failure) and 1 (success) values
    - https://github.com/Feranmi-Alalade/IBM-Data-Science-Capstone-SpaceX/blob/main/3_Data_Wrangling.ipynb

EDA with Data Visualization

Different charts were plotted to make inferences during EDA

- Scatter point charts were plotted to view the relationship between:
  - ‘FlightNumber and Launch Site’,
  - Payload mass and Launch site: It was observed that for the VAFB-SLC launchsites, there are no launches for payload mass greater than 10000kg.
  - FlightNumber and Orbit
- A barplot to show the success rate for each each orbit
- A line chart to show the launch success trend yearly
  - https://github.com/Feranmi-Alalade/IBM-Data-Science-Capstone-SpaceX/blob/main/5_EDA_Viz_lab.ipynb

EDA with SQL

SQL Queries used during the Project:

- Using DISTINCT( ) to select the unique launch sites in the space mission.
    - Select distinct(launch_site) from soacextbl
- To display 5 records where launch sites being with ‘CCA’
    - Select * from spacextbl where launch_site like ‘CCA%’ LIMIT 5
- Display the total payload mass carried by boosters launches by NASA (CRS)
    - Select customer, sum(payload_mass__kg) as sum fromspacextbl

      where customer like ‘NASA (CRS)’

      group by customer
- Display average payload mass carried by booster version F9 v1.1
    - select avg(payload_mass__kg_) from spacextbl where booster_version = 'F9 v1.1'

EDA with SQL

SQL Queries used during the Project:

- List the date when the first successful landing outcome in ground pad was acheived.
    - select date from spacextbl where landing_outcome like 'Success (ground pad)%' and date = (select min(date) from spacextbl)
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
    - select booster_version from spacextbl where payload_mass__kg_ between 4000 and 6000 and landing_outcome like 'Success%drone'
    - https://github.com/Feranmi-Alalade/IBM-Data-Science-Capstone-SpaceX/blob/main/EDA_SQL.ipynb

Build an Interactive Map with Folium

Interactive Map Stages

- The Launch Sites were first located on the map using markers (to mark the different coordinates on the map).
- The different launches from each launch site were then clustered together (using MarkerCluster) to identify them as one, and an icon was added to label the launches that were successful (Green) and unsuccessful (Red).
- Polylines were used to mark the distances from the launch sites to the nearest coastline, city and railway line.
- Images are shown in the next slide.

- https://github.com/Feranmi-Alalade/IBM-Data-Science-Capstone-SpaceX

Folium

Markers showing the different Launch Sites

Folium

Marker Clusters showing the different Launch Sites

Folium

Icons showing the success rate

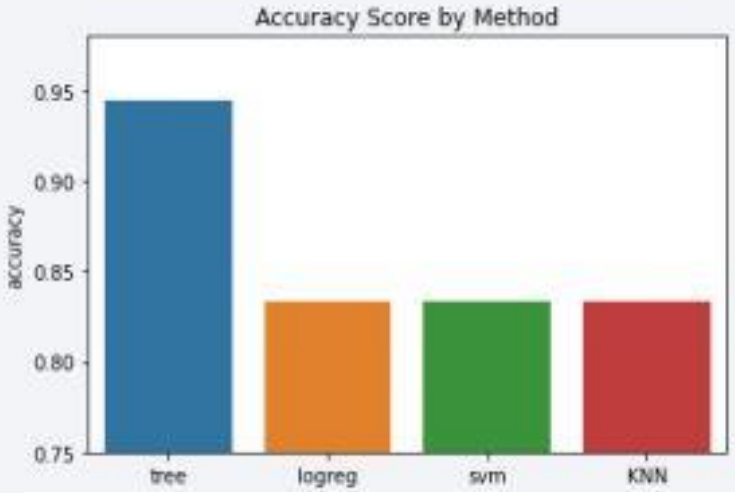
Folium

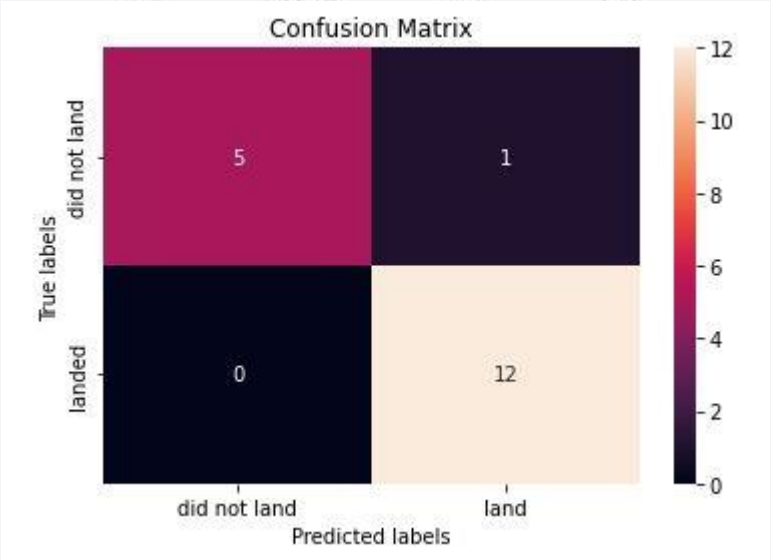
Polylines showing the distance to the coastline

Predictive Analysis (Classification)

- In this project, we predicted the landing outcome of the Falcon 9 first stage using Data Science methodology and Machine Learning algorithms. SpaceX rocket launches costs about 65 million dollars while other providers cost about 165 million upwards. This is because SpaceX reuses its first stage, which is also called the booster stage. The booster returns to the earth surface landing on either ground pads, ocean ships or Drone ships. This project is based on previous launch data (the landing outcomes of the boosters of previous launches), and algorithm predicts the landing outcomes of future launches.
- A new binary column ‘Class’ was added to the dataframe, 0 for unsuccessful landing and 1 for successful landing.
- Machine Learning algorithms namely Support Vector Machines, K-Nearest Neighbors, Logistic Regression and Decision trees were then trained on the features: Orbit, PayloadMass, Flights, Block, etc.
- The algorithms were then used to predict the landing of the rockets

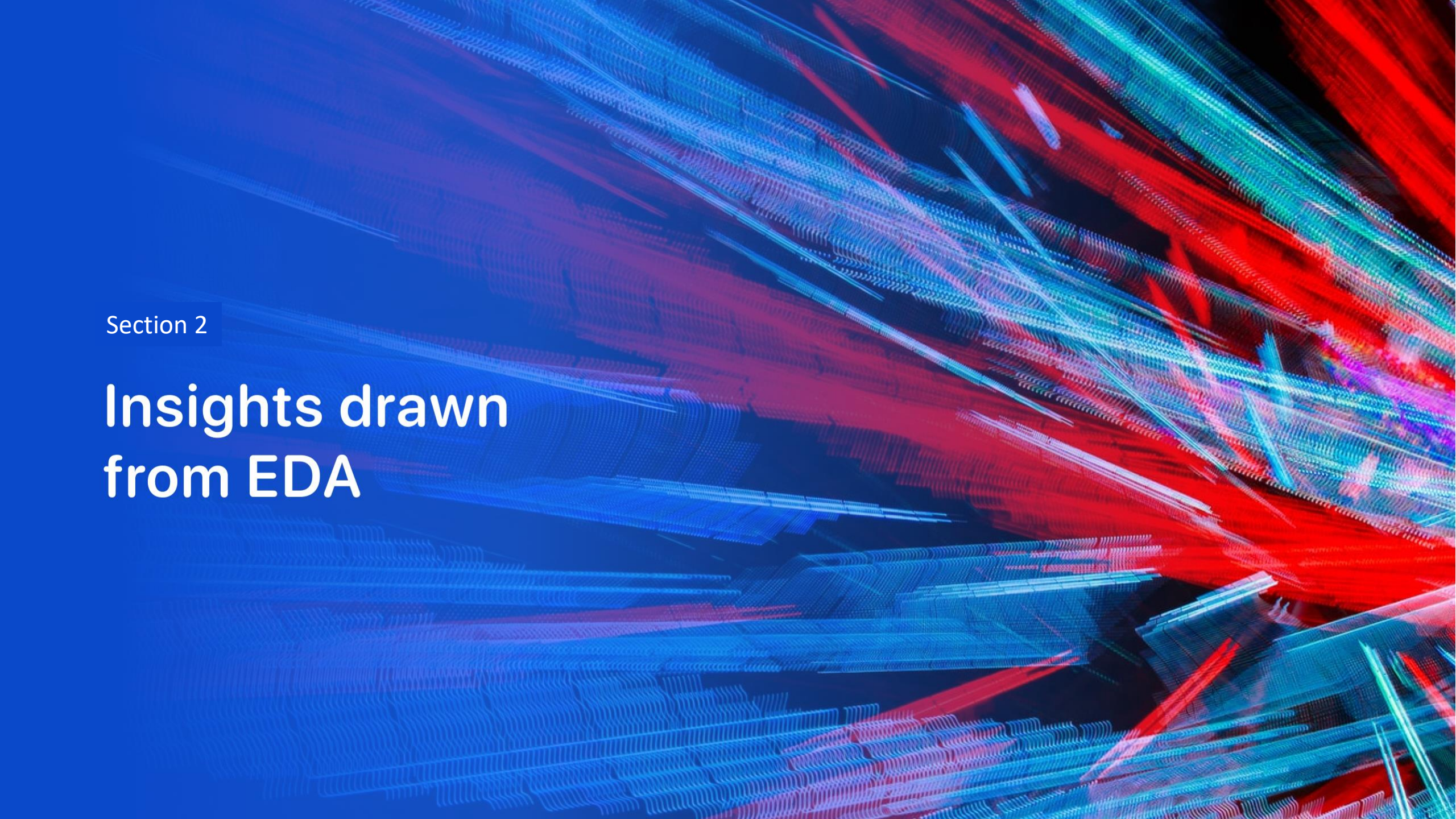
Results



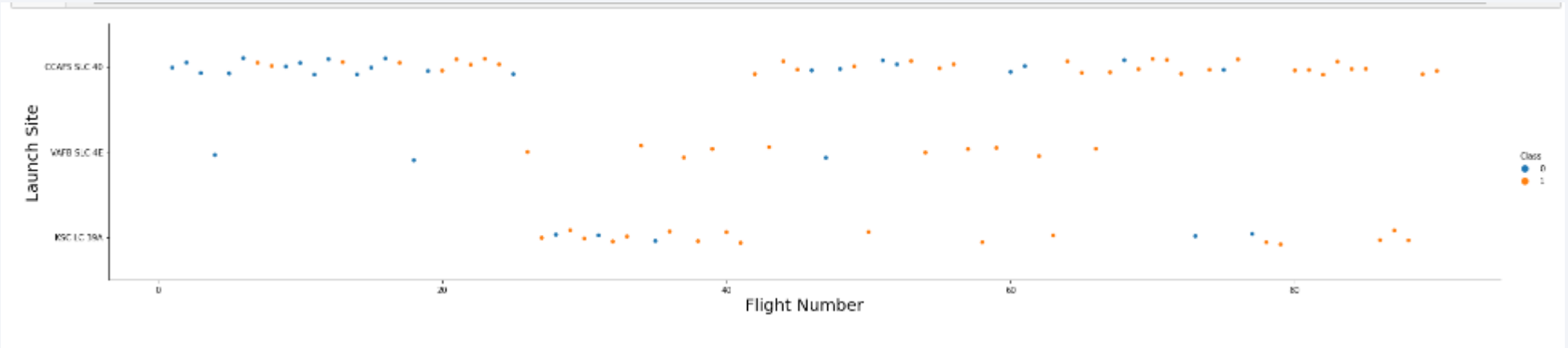


The Decision tree model was the model with the highest accuracy of almost 95%
The confusion matrix is also added in this slide

https://github.com/Feranmi-Alalade/IBM-Data-Science-Capstone-SpaceX/blob/main/7_Machine_learning_Prediction.ipynb

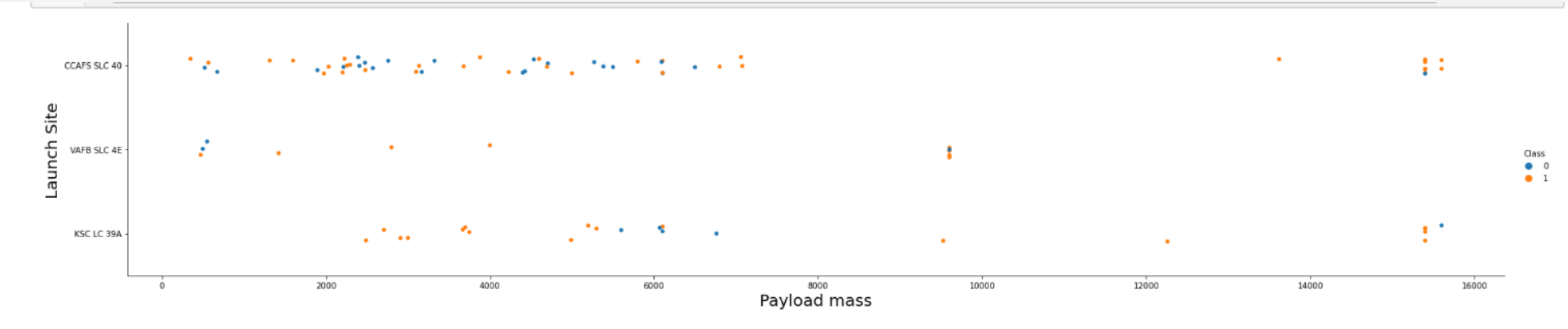

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



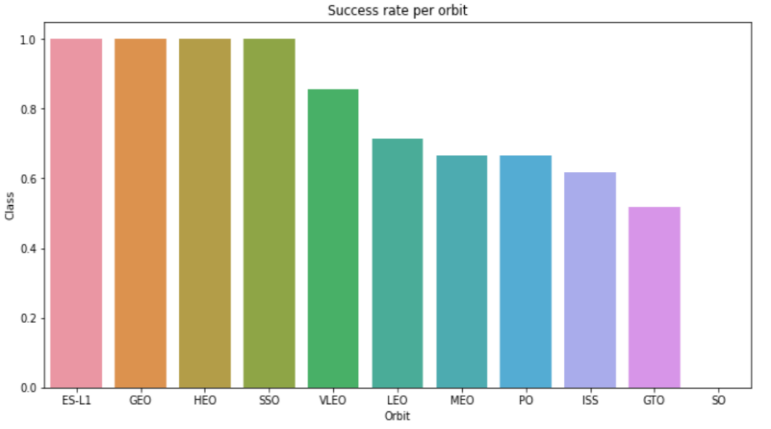
We can see in this chart that the earlier launches took place at the CCAFS SLC-40 launch site, we can also see that the CCAFS SLC-40 is the preferred location for launches.

Payload vs. Launch Site

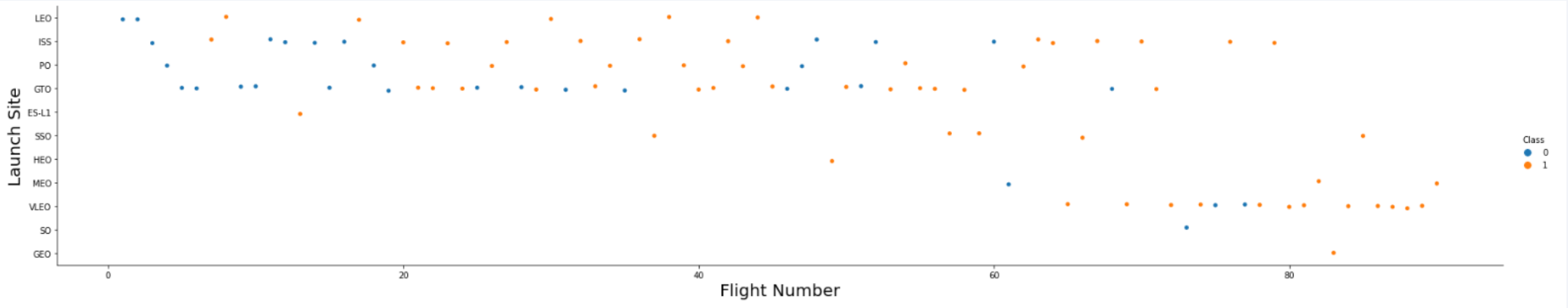


We can see in this chart that launches with payload mass of above 10,000kg do took place at the CCAFS SLC-40 launch site, we can also see that launches with payload mass of less than about 2500kg do not take place at KSC LC 39A launch site.

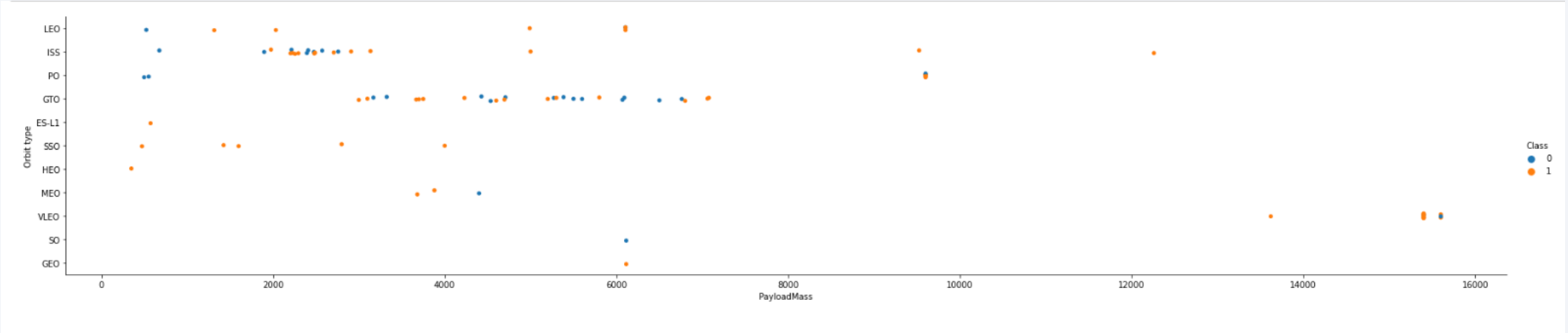
Success Rate vs. Orbit Type



Launches to GTO orbit are clear underdogs with lowest success rate (among orbits with at least 3 missions) with visually small gain over time. This a high geosynchronous orbit located at 22,236 km above Erth. Contrary to it, low-distance orbits (VLEO, ISS, LEO) show much better results

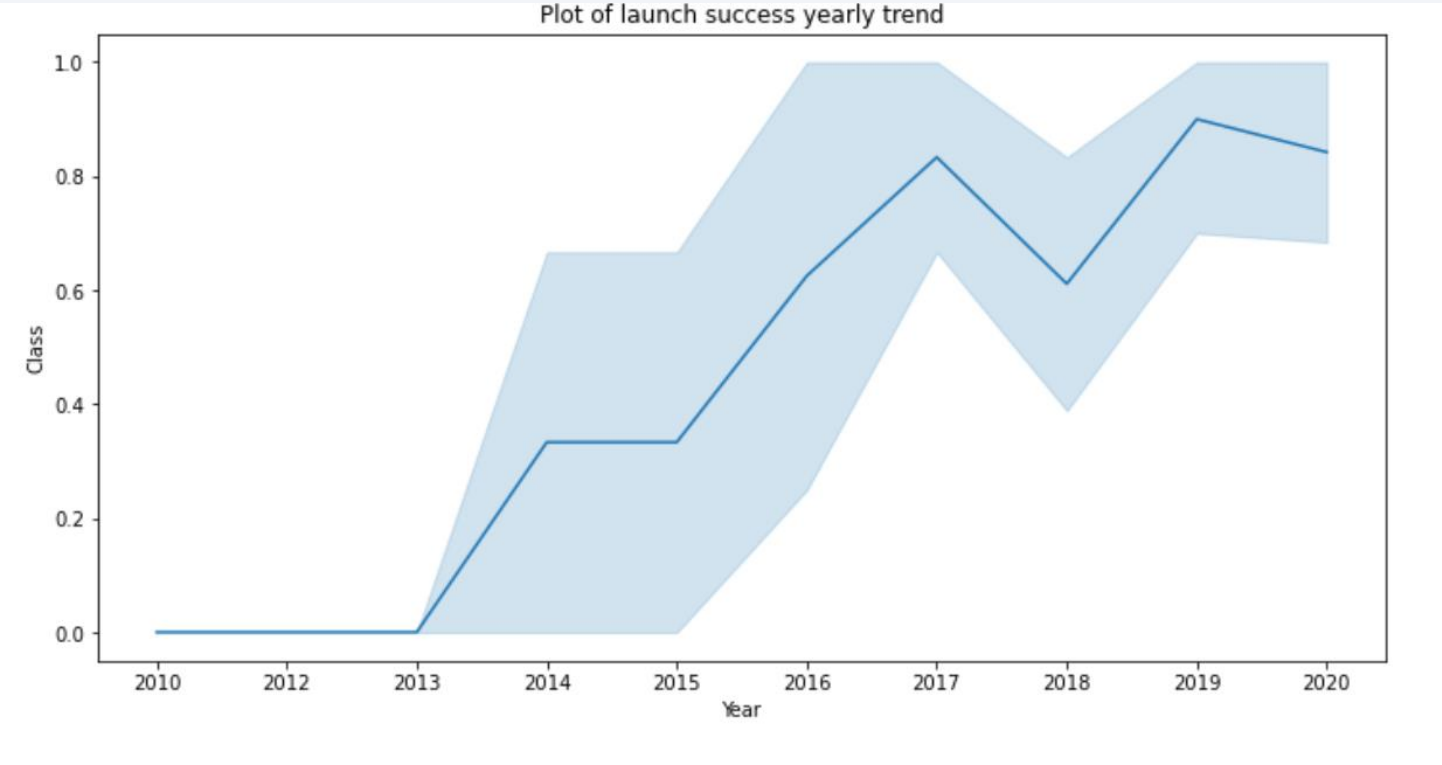


Payload vs. Orbit Type



All of the medium and high payload launches happen to low-Earth orbits (VLEO, ISS, PO), with high success rate. Lower payloads can be observed both at geosynchronous orbits and low-Earth orbits, generally, with significantly lower rate of successful landings.

Launch Success Yearly Trend



The Success rate increases yearly due to improved technology and experience.

All Launch Site Names

```
1 %sql select distinct(launch_site) from spacextbl
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5b-8580-3fbfa4
Done.
```

5]:

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

Though 4 sites are present in initial database, actually all launches were performed from 3 sites – CCAFS LC-40 was renamed into CCAFS SLC-40 and means “Cape Canaveral Space Launch Complex”, Florida. Nearby located KSC LC-39A which i.e. “Kennedy Space Center Launch Complex”. VAFB SLC-4E is in California and the full name is “Vandenberg Space Launch Complex”.

Launch Site Names Begin with 'CCA'

```
1 %sql select * from spacextbl where launch_site like 'CCA%'
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31505/BLU
DB
Done.
```

[6]:

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

These are the first five launches that took place at launch sites starting with ‘CCA’

Total Payload Mass

```
%sql select customer, sum(payload_mass__kg_) as sum from spacextbl\
where customer like 'NASA (CRS)'\
group by customer
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08
Done.
```

customer	SUM
NASA (CRS)	45596

The total payload mass of all launches that belonged to the NASA (CRS) was 45596kg

Average Payload Mass by F9 v1.1

```
1 %sql select avg(payload_mass__kg_) from spacex
2 where booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5b-858
Done.
```

8]:

1
2928

The average payload mass of all launches that where the booster version was ‘F9 v1.1’ is 2928kg, which is not very much.

First Successful Ground Landing Date

The date of the first successful ground landing date was 2015 which was five years after the first launch in 2010.

```
%sql select min(date) as Date from spacextbl\
where landing_outcome like 'Success (ground pad)'
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5b-8580-3fbf
Done.
```

4]:

DATE
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

```
1 %sql select booster_version from spacextbl\
2 where landing_outcome like 'Success (drone ship)' and payload_mass__kg_ between 4000 and 6000
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1od8lcg.databases.appd
Done.
```

0]:

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

This shows the booster version of the successful drone ship landing with payload between 4000 and 6000kg

Total Number of Successful and Failure Mission Outcomes

```
%sql select count(*) from spacextbl\
where mission_outcome like 'Success%'
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5
Done.
```

]:

1
100

```
%sql select count(*) from spacextbl\
where mission_outcome like 'Fail%'
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5
Done.
```

]:

1
1

All launches according to Mission outcomes except one were successful

Boosters Carried Maximum Payload

```
1 %sql select booster_version from spacextbl\
2 where payload_mass__kg_ = (select max(payload_mass__kg_) from spacextbl)
```

```
 * ibm_db_sa://dfh73072:***@ea286ace-86c7-4d5b-8580-3fbfa46b1c66.bs2io90l08kqb1c
Done.
```

6]:

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

The booster versions where the launches carried the maximum payload mass are listed

2015 Launch Records

landing_outcome	booster_version	launch_site
Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

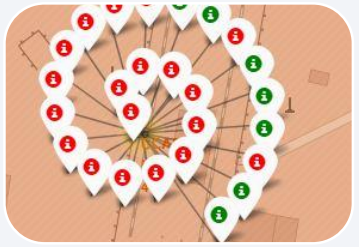
2015 launch records where the booster failed landing on the drone ships

Section 3
Launch Sites
Proximities Analysis

LAUNCH SITE CLUSTERS

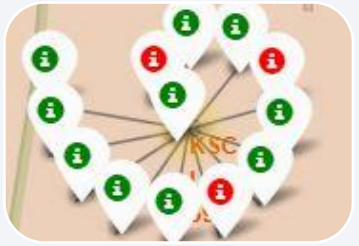
The launches took place in two clusters: 10 in California and 46 in Florida

SUCCESS RATES



CCAFS LC-40

CCAFS SLC-40

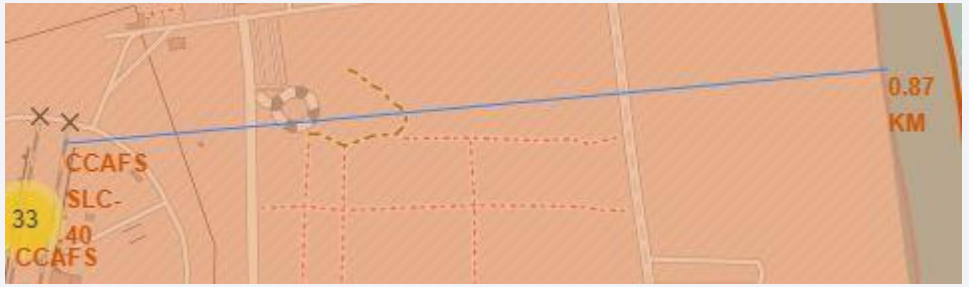


KSC LC-39A

VAFB SLC-4E

Information about the launch sites can be gotten by clicking on the different clusters. KSC LC-39A has the highest success rate.

Proximity

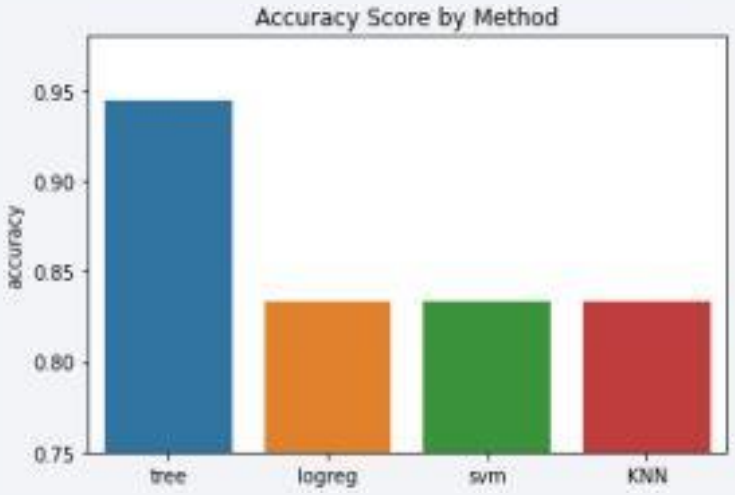


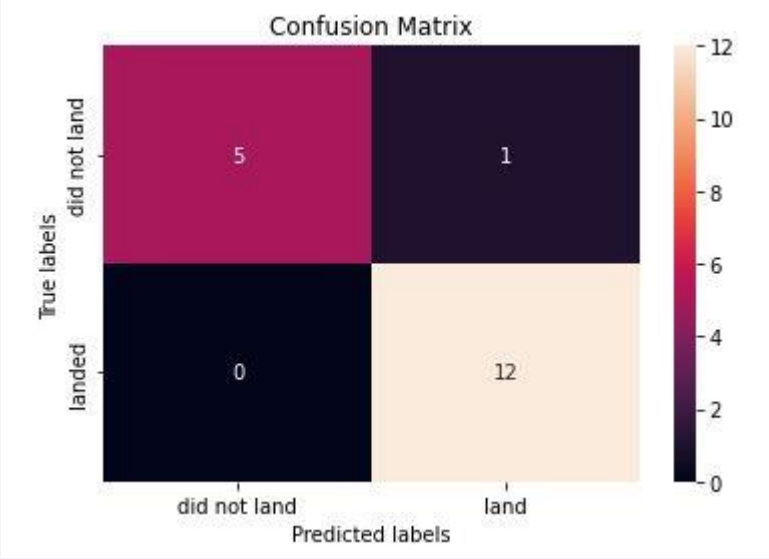
The launch sites are closer to the coast line(0.87km) relative to the city(78km) and the rail line(22km)

Section 5

Predictive Analysis (Classification)

Classification Accuracy





The Decision tree model returned the highest accuracy with an accuracy of 91.4%. The confusion matrix also shows a high True Negative rate and a high True Positive rate

Conclusions

- Four different algorithms were trained on the same data, and the Decision tree model was able to predict 12 out of 13 successful landings and 5 out of 5 unsuccessful landings.

Thank you!