IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mohammed Jamil
2022-08-31

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection through API and Web Scraping
  - Data Wrangling
  - Exploratory Data Analysis with SQL and Data Visualization
  - Interactive Dashboarding with Plotly/Dash
  - Machine Learning classification

- Summary of all results
  - Valuable data was collected through SpaceX API and Web scraping.
  - Exploratory Data Analysis with SQL and Data Visualization allowed to identify the features that are best for building the ML model.
  - Different classification models were built to identify which one performs the best.

Introduction

- Project background and context :

  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 Space X 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. This goal of the project is to create a machine learning pipeline to predict if the first stage will land successfully.

- Desirable answer :
  - What factors determine the successful landing of the Falcon 9 first stage ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Data was obtained using SpaceX API and Web scraping the wikipedia page related to Falcon rocket launches.
- Perform data wrangling
  - Categorical variables were transformed using one-hot-encoding including landing outcome.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - After the data was collected, it was normalized and divided into training and testing sets and evaluated by different classification algorithms.

Data Collection

Data about SpaceX Falcon 9 launches was collected from 2 sources:

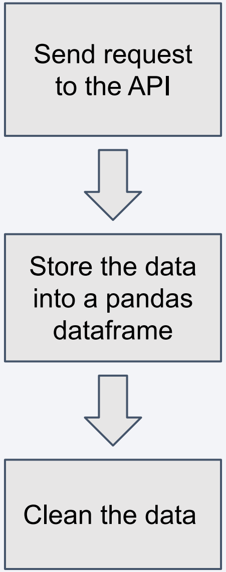
- SpaceX API.
- Web scraping the wikipedia page.

Data Collection – SpaceX API

- SpaceX provides a public API that was used to collect data about rocket launches.

- File link :

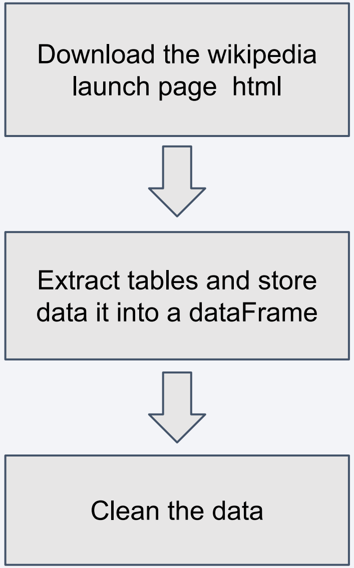
  https://github.com/MohammedJamil/IBM-applied-data-science-capstone/blob/master/Data-collection.ipynb



Data Collection - Scraping

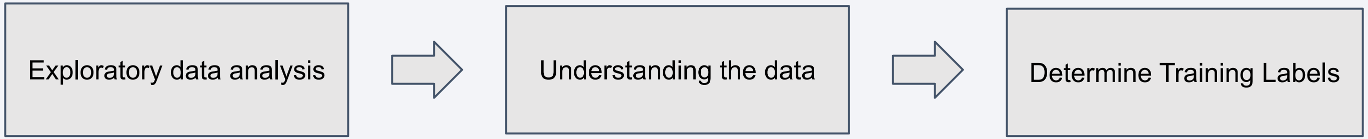
- Wikipedia tables are also a great source of data. Web scraping using BeautifulSoup allows to gets this valuable data.

- File link : https://github.com/MohammedJamil/IBM-applied-data-science-capstone/blob/master/Data-collection-with-web-scraping.ipynb



Data Wrangling

- Performing exploratory data analysis on the dataset helped finding patterns in the data and determine what would be the label for training supervised models by calculating the number and occurrence of each orbit and the number and occurence of mission outcome per orbit type and finally, creating a landing outcome label.

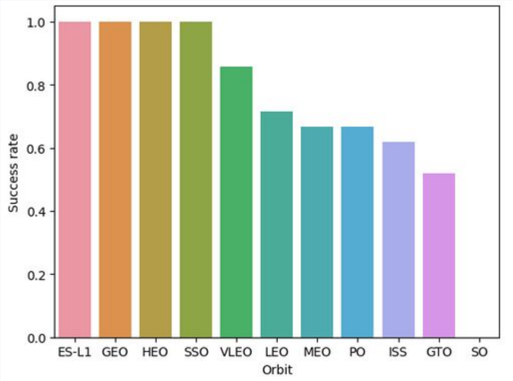


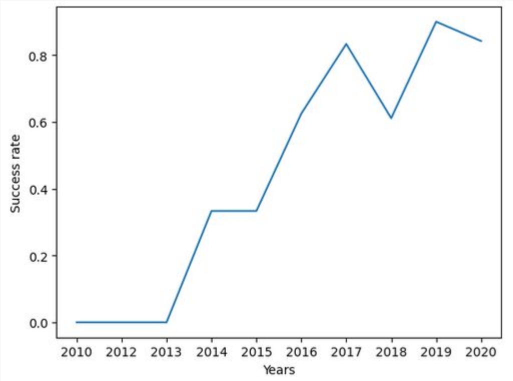
- File link :

  https://github.com/MohammedJamil/IBM-applied-data-science-capstone/blob/master/Data-wrangling.ipynb

EDA with Data Visualization

- EDA with visualizing helped us exploring the relationship between flight number and launch Site, payload and launch site, success rate of each orbit type, flight number and orbit type, the launch success yearly trend.





- FIle link :

  https://github.com/MohammedJamil/IBM-applied-data-science-capstone/blob/master/EDA-data-visualization.ipynb

EDA with SQL

- The following SQL queries were performed to:
  - **Display the names of the unique launch sites in the space mission**
  - **Display 5 records where launch sites begin with the string 'CCA'**
  - **Display the total payload mass carried by boosters launched by NASA (CRS)**
  - **Display average payload mass carried by booster version F9 v1.1**
  - **List the date when the first successful landing outcome in ground pad was achieved.**
  - **List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000**
  - **List the total number of successful and failure mission outcomes**
  - **List the names of the booster_versions which have carried the maximum payload mass. Use a subquery**
  - **List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015**
  - **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**
- File link :

  https://github.com/MohammedJamil/IBM-applied-data-science-capstone/blob/master/EDA-sql.ipynb

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers indicate points like launch sites.
  - Circles indicate highlighted areas around specific coordinates, like NASA Johnson Space Center.
  - Marker clusters indicates groups of events in each coordinate, like launches in a launch site and.
  - Lines are used to indicate distances between two coordinates.

- File link :

  https://github.com/MohammedJamil/IBM-applied-data-science-capstone/blob/master/Data-visualization-folium.ipynb

Build a Dashboard with Plotly Dash

- The following graphs and plots were used to visualize data
  - Pie charts showing the total launches by a certain sites
  - Scatter graph of the relationship between Outcome and Payload Mass for the different booster version.

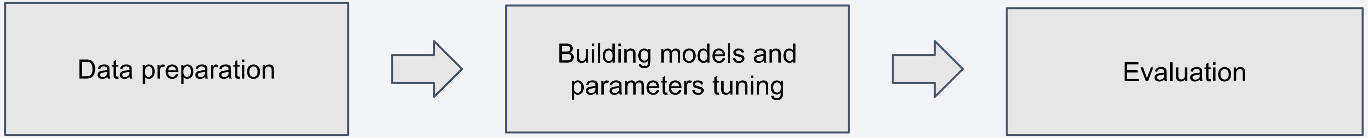
- This combination allowed to quickly analyze the relation between payloads and launch sites, helping to identify where is best place to launch according to payloads.

- File link:

  https://github.com/MohammedJamil/IBM-applied-data-science-capstone/blob/master/spacex_dash_app.py

Predictive Analysis (Classification)

- Data was loaded using numpy and pandas, then transformed and splited into training and testing. Different machine models learning were built and tuned using different hyperparameters using GridSearchCV. Then the different models were evaluated.



- File link:

  https://github.com/MohammedJamil/IBM-applied-data-science-capstone/blob/master/Predictive-analysis.ipynb

Results

- The SVM, KNN, and Logistic Regression models are the best in terms of prediction accuracy 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.

- KSC LC 39A had the most successful launches from all the sites.

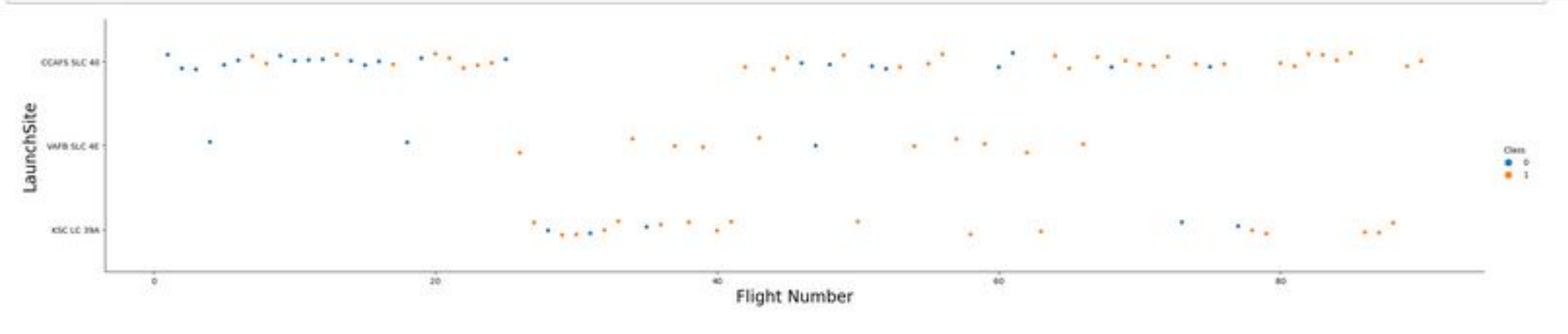
- Orbit GEO,HEO,SSO,ES L1 has the best Success Rate.

Section
2

Insights drawn from EDA

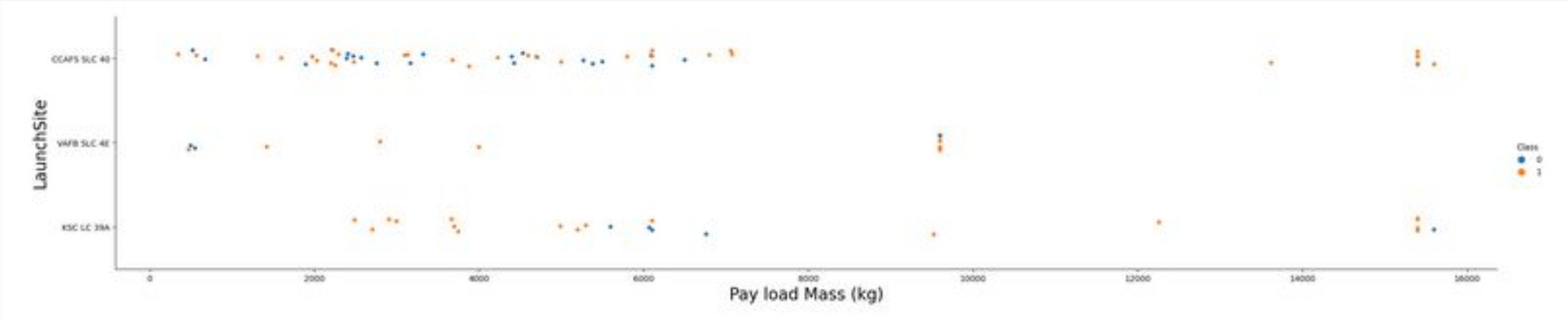
Flight Number vs. Launch Site

- Launches from the site of CCAFS SLC 40 are significantly higher than launches form other sites.



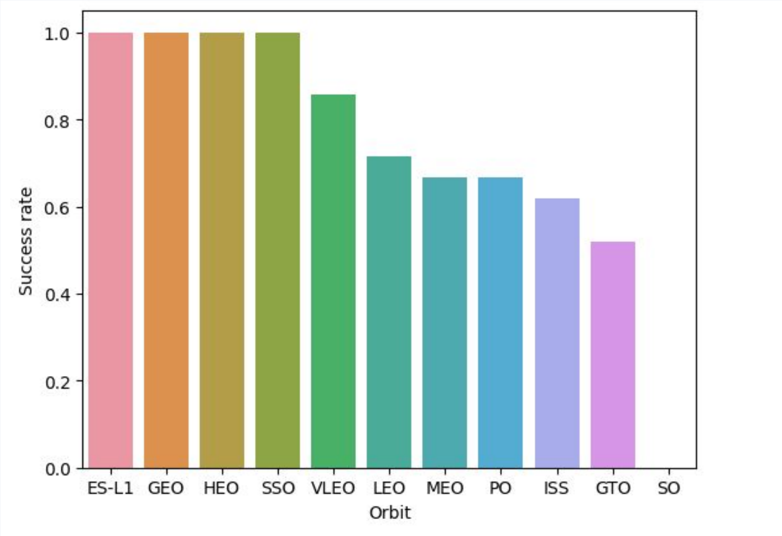
Payload vs. Launch Site

- Payloads over 9,000kg have excellent success rate.
- Payloads over 12,000kg seems to be possible only on CCAFS SLC 40 and KSC LC 39A launch sites.



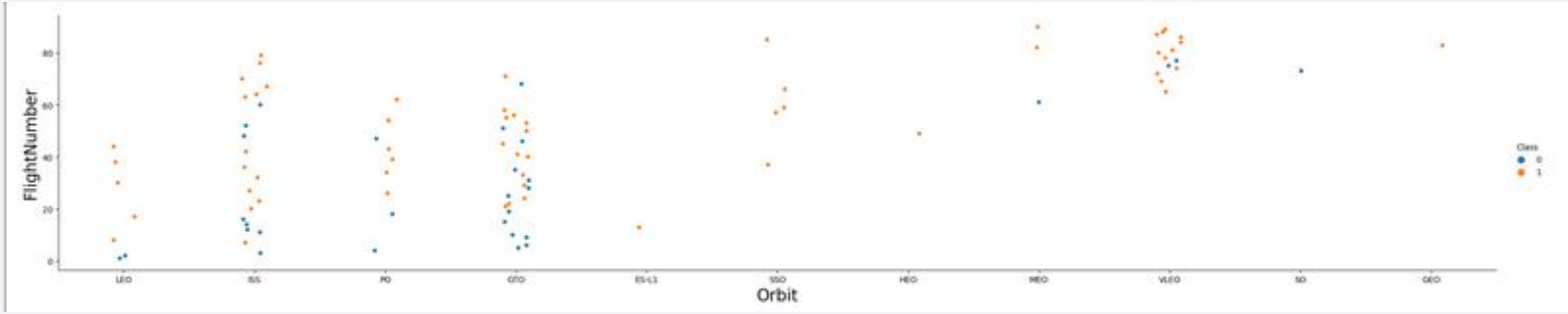
Success Rate vs. Orbit Type

- The orbit types of ES-L1, GEO, HEO, SSO are among the highest success rate.



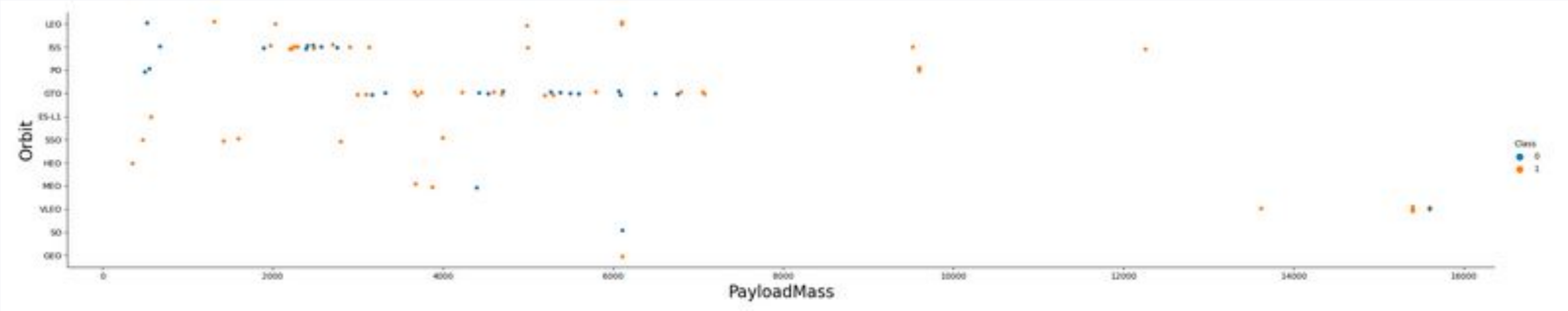
Flight Number vs. Orbit Type

- For the LEO orbit the Success appears related to the number of flights; on the other hand, there seems to be no relationship between flight number when in GTO orbit.



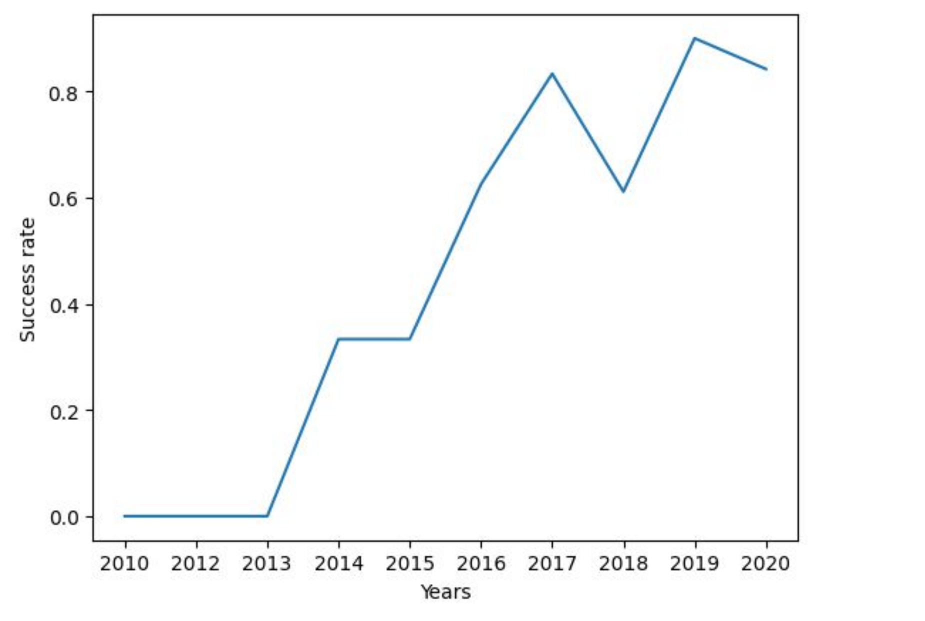
Payload vs. Orbit Type

- With heavy payloads the successful landing or positive landing rate are more for Polar,LEO and ISS.
- However for GTO we cannot distinguish this well as both positive landing rate and negative landing(unsuccessful mission) are both there here.



Launch Success Yearly Trend

- you can observe that the success rate since 2013 kept increasing till 2020



All Launch Site Names

- The keyword is used DISTINCT to show only unique launch sites from the SpaceX data.

```
%%sql

select distinct(LAUNCH_SITE) from SPACEXTBL
```

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

Launch Site Names Begin with 'CCA'

- The query displays 5 records where launch sites begin with `CCA`

```
%%sql

selec * from SPACEXTBL
where LAUNCH_SITE like 'CCA%'
limit 5
```

```
 * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databa
ses.appdomain.cloud:32731/bludb
Done.
```

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	land
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Fail
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	Fail
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	

Total Payload Mass

- We calculated the total payload carried by boosters from NASA as 45596 using the query below

```
In [8]: %%sql

        select sum(PAYLOAD_MASS__KG_) from SPACEXTBL
        where CUSTOMER = 'NASA (CRS)'

         * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databa
        ses.appdomain.cloud:32731/bludb
        Done.
```

Out[8]:

1
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is be calculated with the query below

```
In [9]: %%sql

        select avg(PAYLOAD_MASS__KG_) from SPACEXTBL
        where BOOSTER_VERSION = 'F9 v1.1'

         * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databa
        ses.appdomain.cloud:32731/bludb
        Done.

Out[9]:
```

1
2928

First Successful Ground Landing Date

- First successful ground landing date is be calculated with the query below :

In [14]:
```
%%sql

select min(DATE) from SPACEXTBL
where LANDING__OUTCOME = 'Success (ground pad)'
```

```
 * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[14]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- The booster version used in the successful drone ship landing with Payload between 4000 and 6000 is be calculated with the query below

```
In [15]: %%sql

         select BOOSTER_VERSION from SPACEXTBL
         where LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ between 4000 and 6000

 * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databa
ses.appdomain.cloud:32731/bludb
Done.
```

Out[15]:

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

Total Number of Successful and Failure Mission Outcomes

- We used wildcard like ‘%’ to filter for WHERE MissionOutcome was a success or a failure.

In [19]:
```
%%sql

select distinct(MISSION_OUTCOME) as OUTCOME, count(*) as TOTAL from SPACEXTBL
group by MISSION_OUTCOME
```

```
 * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databa
ses.appdomain.cloud:32731/bludb
Done.
```

Out[19]:

outcome	total
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- Boosters versions that carried maximum payload can be calculated with the query below

```
In [24]: %%sql

select BOOSTER_VERSION from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select MAX(PAYLOAD_MASS__KG_) from SPACEXTBL)

 * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[24]:

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

- We determined the booster that have carried the maximum payload using a subquery in the WHERE clause and the MAX() function.

```
In [28]: %%sql

select LANDING__OUTCOME, BOOSTER_VERSION, LAUNCH_SITE from SPACEXTBL
where LANDING__OUTCOME = 'Failure (drone ship)' and date_part('year', DATE) = 2015

 * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databa
ses.appdomain.cloud:32731/bludb
Done.
```

Out[28]:

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

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

- Rank Landing Outcomes Between 2010-06-04 and 2017-03-20 can be calculated with the query below

```
In [32]: %%sql

select distinct(LANDING__OUTCOME), count(*) as TOTAL from SPACEXTBL
where DATE between '2010-06-04' and '2017-03-20'
group by LANDING__OUTCOME
order by TOTAL DESC
```

```
 * ibm_db_sa://sqz24887:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databa
ses.appdomain.cloud:32731/bludb
Done.
```

Out[32]:

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

Section
3

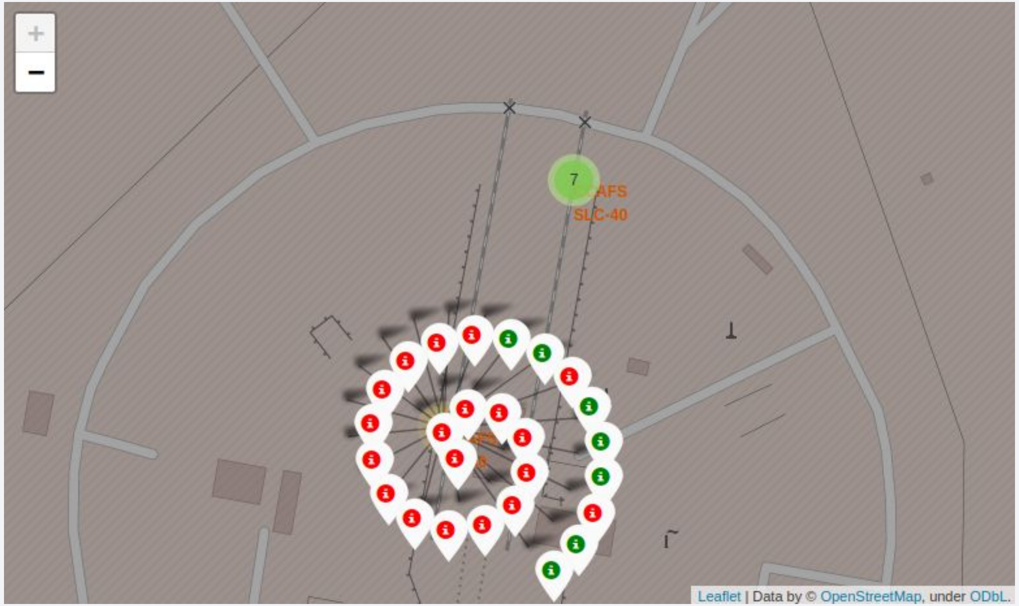
Launch Sites Proximities Analysis

All launch sites

- Launch sites are near sea, probably by safety, but not too far from roads and railroads

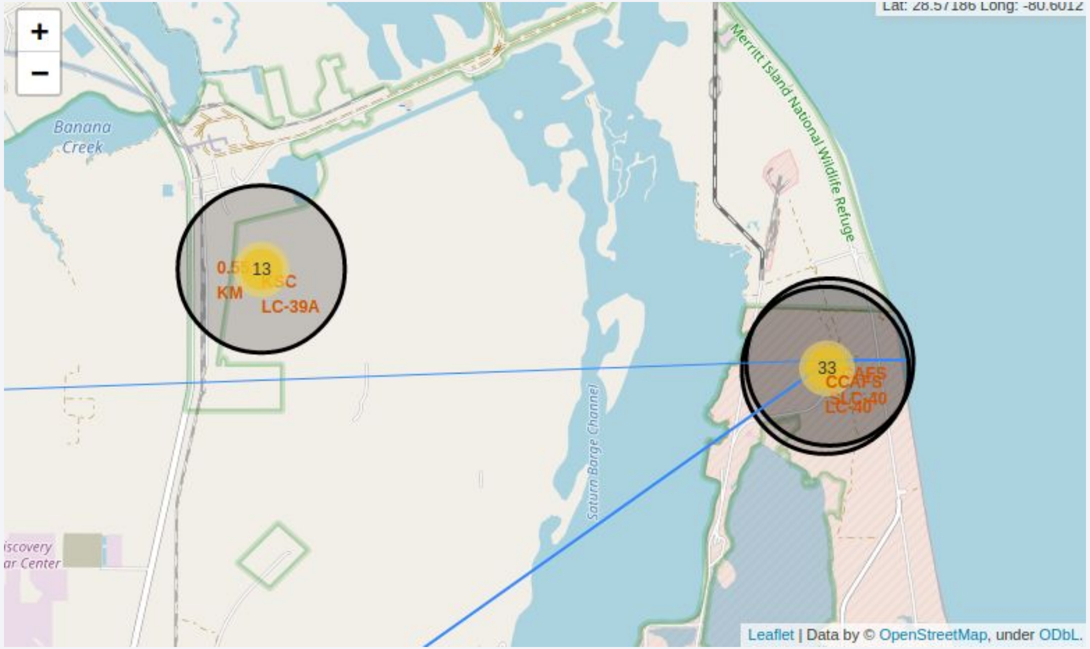
Launch Outcomes by Site

- Green markers indicate successful and red ones indicate failure.



Logistics and Safety

- Launch site has good logistics aspects, being near railroad and road and relatively far from inhabited areas.

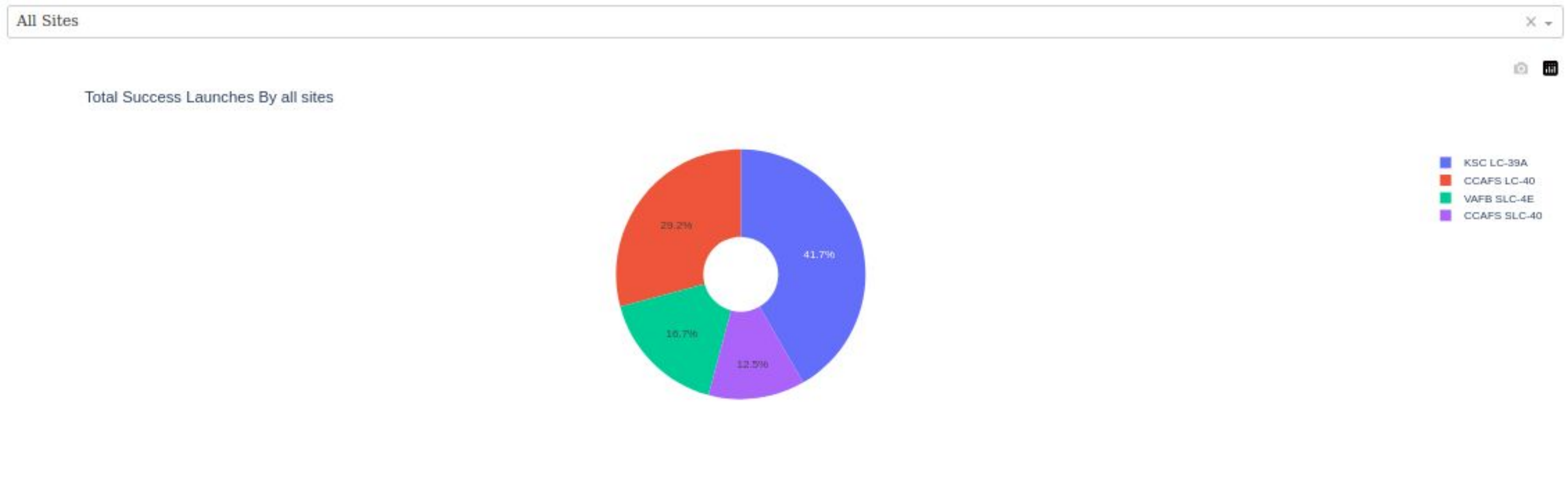


Section
4

Build a Dashboard with Plotly Dash

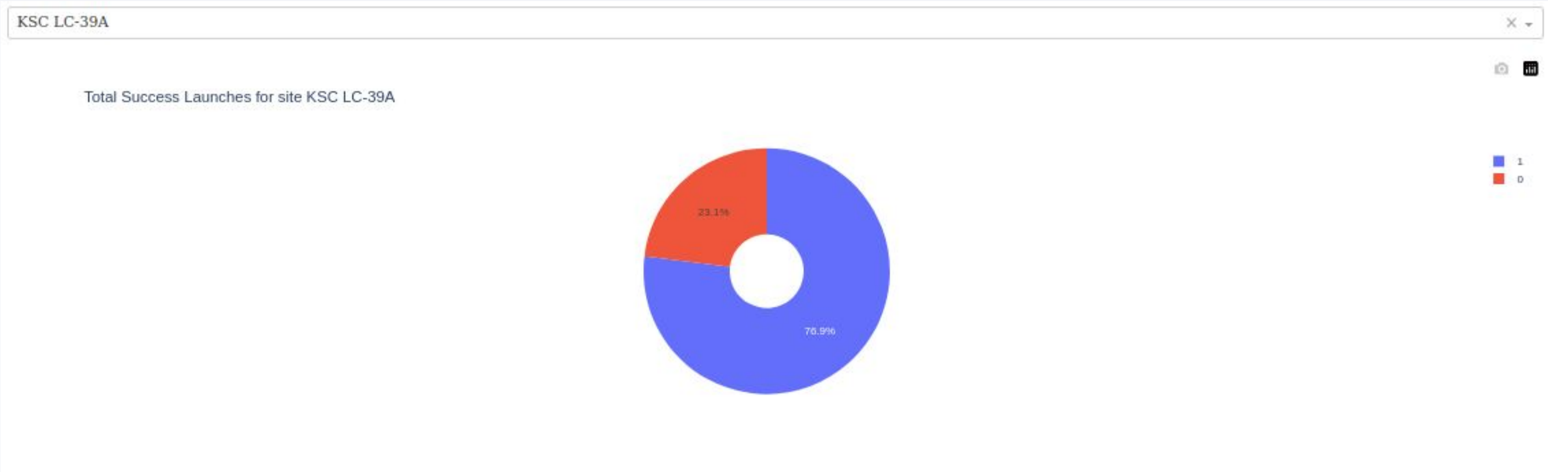
Successful Launches by Site

- KSC LC-39A have the highest number o successful launches followed by CCAFS LC-40



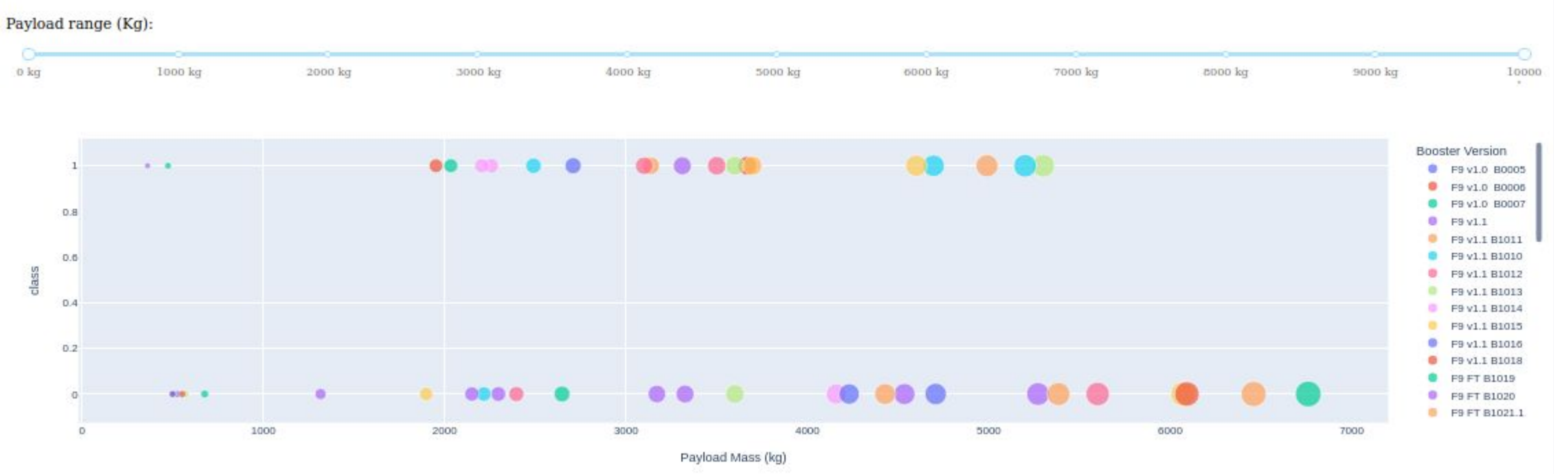
Launch Success Ratio for KSC LC-39A

- The success ratio of KSC LC-39A is 76.9%



Payload vs. Launch Outcome

- Payloads above 6,000kg are the least successful.

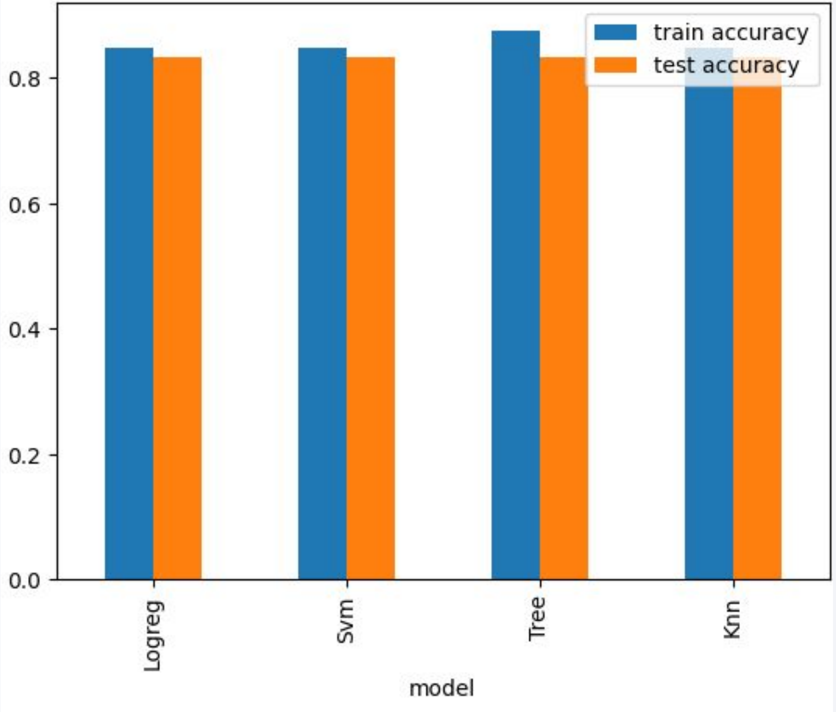


Section
5

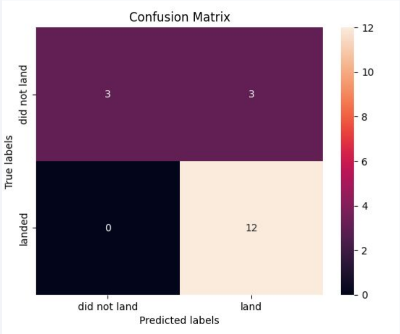
Predictive Analysis (Classification)

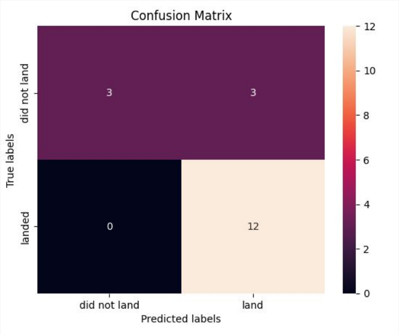
Classification Accuracy

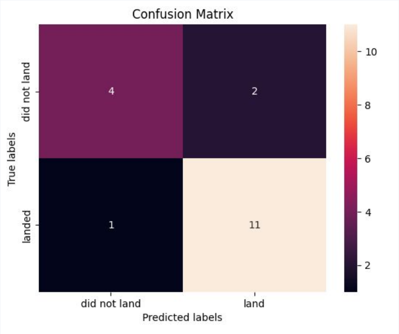
- All models are the same accuracy on the testing dataset, while decision tree model have the highest training accuracy

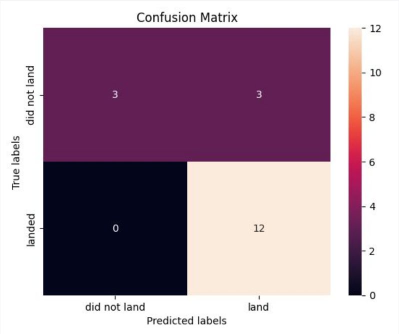


Confusion Matrix









Conclusions

- The best launch site is KSC LC-39A;
- Launches below 6,000kg are less risky;
- Most of mission outcomes are successful, successful landing outcomes seem to improve over time.
- Decision Tree Classifier can be used to predict successful landings and increase profits.

Thank you!