IBM Data Science Professional Certificate

Applied Data Science Capstone Project

Mustafa Türköz

https://github.com/mustafaturkoz

29.11.2021

Outline

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

Executive Summary

- SpaceX Data are collected from public SpaceX API and SpaceX Wikipedia page. New column, labelled as ‘class’ are created which displays successful landings. Data are explored by using SQL, visualization, folium maps, and dashboards. Relevant columns are used as features. All categorical variables are changed via one hot encoding.
- Four machine learning models are implemented. These are Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. The result of these models are similar.

Introduction

- The commercial space age is here, companies are making space travel affordable for everyone. There are some companies such as Virgin Galactic, Rocket Lab and Blue Origin. Yet, the best most successful one is SpaceX. The reason behind SpaceX success is relatively inexpensive. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other manufacturers' cost are up to 165 million dollars. There are three stages in regular rockets. Recovering stage of rockets aids reduction which is valid for SpaceX. Space X recovers the first stage.
- Space Y aims to compete with SpaceX. Space Y tries to predict successful recovering Stage 1. Now, here we are using machine algorithms to train SpaceX data in order to fulling demands of Space Y as data science team.

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX Data are collected from public SpaceX API and SpaceX Wikipedia page Perform data wrangling
  - New column labelled as ‘class’ are created which displays true successful landings.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models via GridSearchCV.

Data Collection

Data Collection process is a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

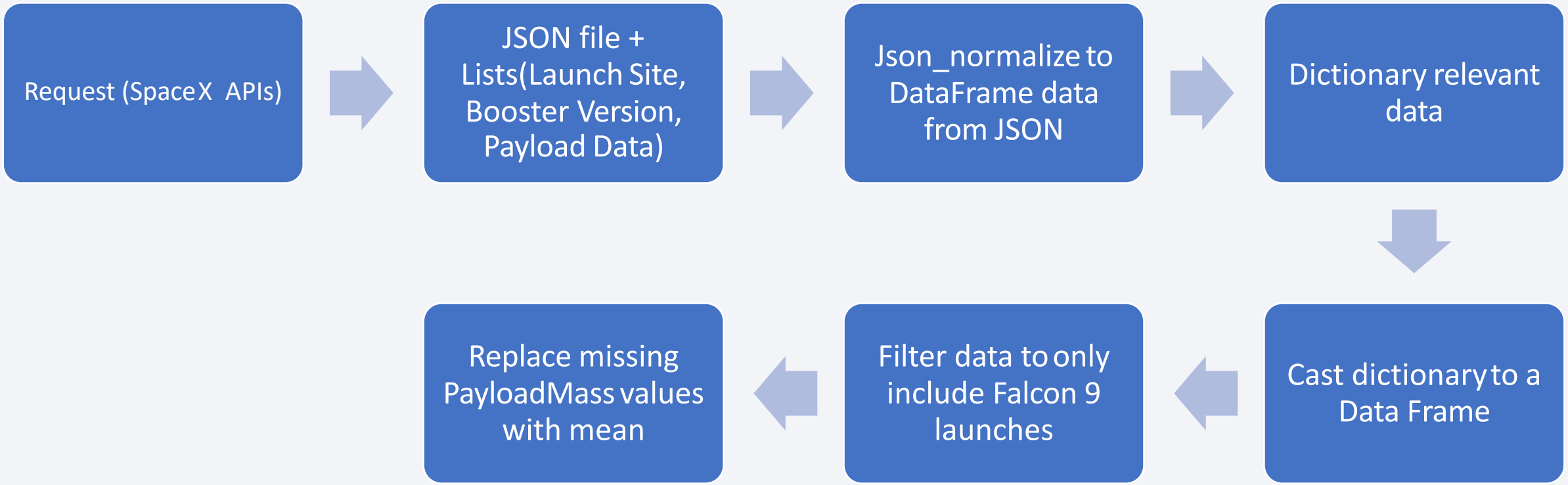
SpaceX API Data Columns:

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Ldegs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Webscrape Data Columns:

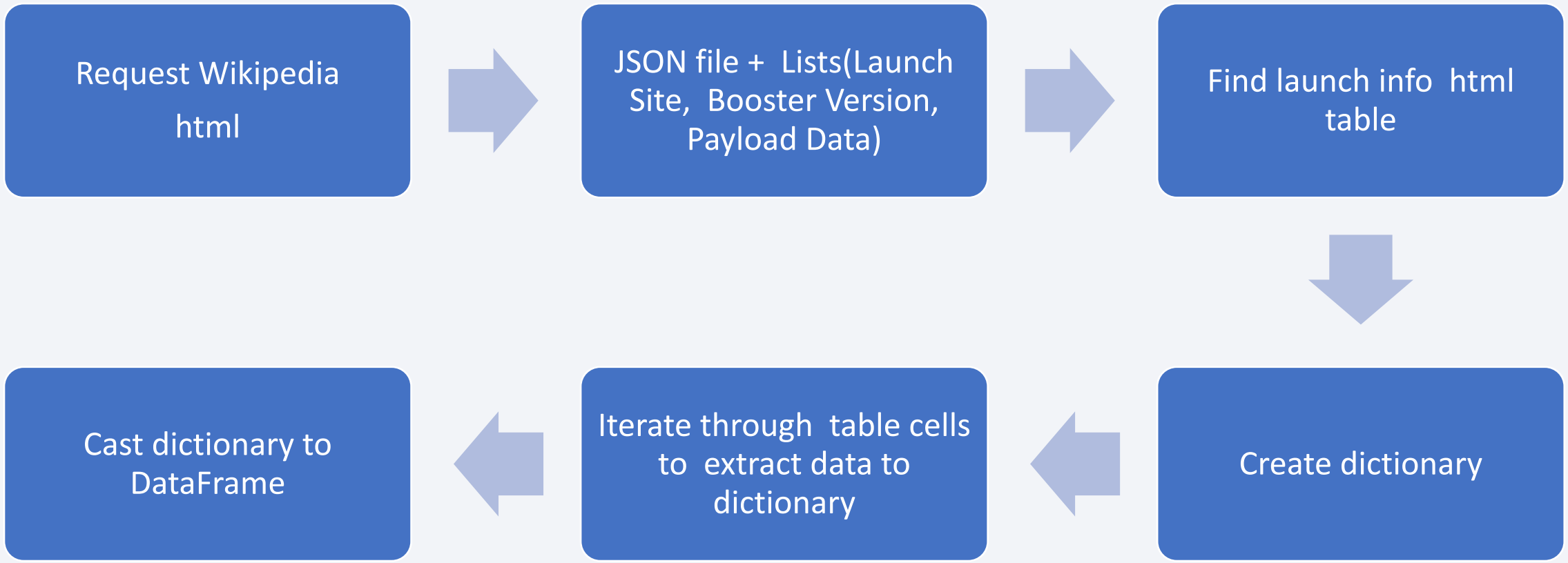
Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection - SpaceX API



Github link:
https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate/blob/main/Week_1/jupyter-labs-spacex-data-collection-api.ipynb

Data Collection - Web Scrapping



Github link:
https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate/blob/main/Week_1/jupyter-labs-webscraping.ipynb

Data Wrangling

- Construct new a label to classify landing outcomes where successful equals to 1 and failure equals to 0.
- Outcome column has two components: ‘Mission Outcome’ and ‘Landing Location’.
- New column ‘class’ has a value 1 when ‘Mission Outcome’ is True else 0 .
- A value mapping as follows:
  - If True ASDS, True RTLS, & True Ocean then 1
  - If None None, False ASDS, None ASDS, False Ocean, False RTLS then 0

Github link:
https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate/blob/main/Week_1/labs-jupyter-spacex-Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis is executed on certain variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.
- Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend are examined.
- In order to investigate relationships of those variables to make decision whether they are informative or not for training machine learning models are scatter plots, line charts and bar plots.

Github link:
https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate/blob/main/Week_2/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- Dataset is loaded into IBM DB2 database.
- Via SQL Python integration, queries are implemented.
- Queries are performed in order to comprehend dataset.
- Queries produces information concerning about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

Github link:
https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate/blob/main/Week_2/jupyter-labs-eda-dataviz.ipynb

Build an Interactive Map with Folium

- Launch Sites, successful and unsuccessful landings, and a proximity example to major locations: Railway, Highway, Coast, and City are indicated by Folium Map.
- Folium Maps mentioned above provides information about Launch Sites location. Moreover, Folium Maps visualize successful landings with respect to location.

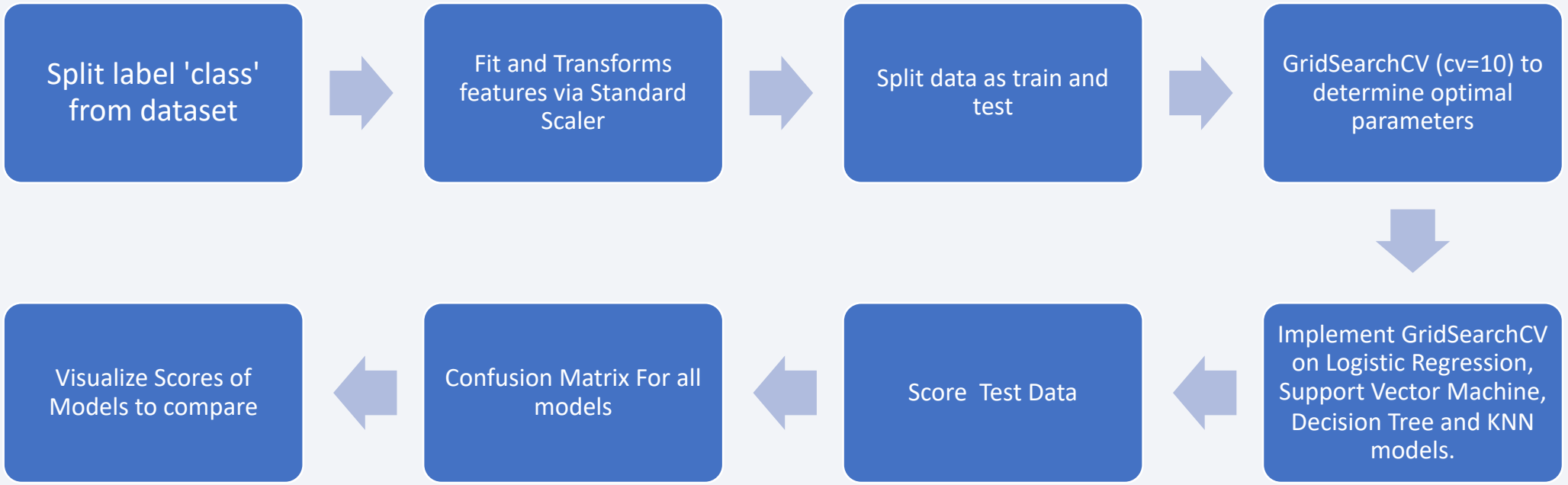
Github link:
https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate/blob/main/_Week_3/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie Chart and Scatter Plot are located in Dashboard.
- In order to display individual launch site success rates, pie chart is picked.
- Scatter plot gets two inputs:
  - All sites or individual site
  - Payload mass on a slider between 0 and 10000 kg
- The scatter plot provides us information how success varies across launch sites, payload mass and booster version.

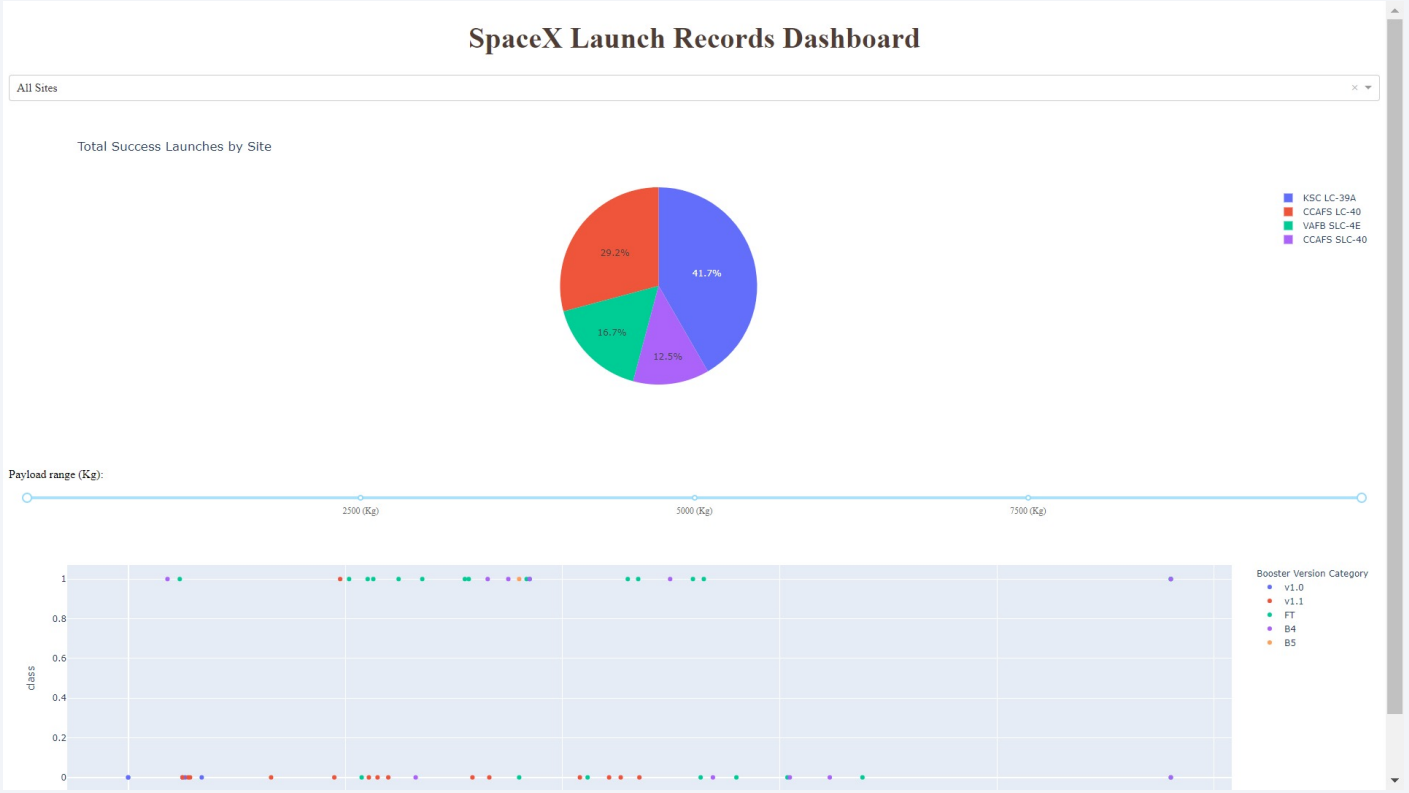
Github link:
https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate/blob/main/_Week_3/SpaceX_Dash_App.py

Predictive Analysis (Classification)



Github link:
https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate/blob/main/_Week_4/SpaceX_Machine%20Learning%20Prediction_Part_5.ipynb

Results

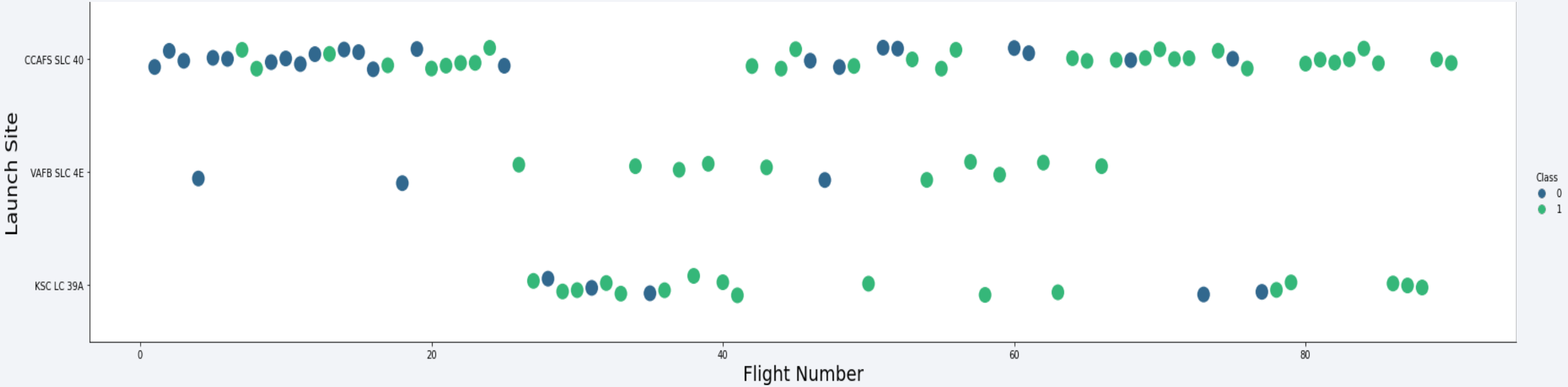


Preview of the Plotly Dashboard

EDA with Visualization

EXPLOTARY DATA ANALYSIS with SEABORN PLOTS

Flight Number vs. Launch Site



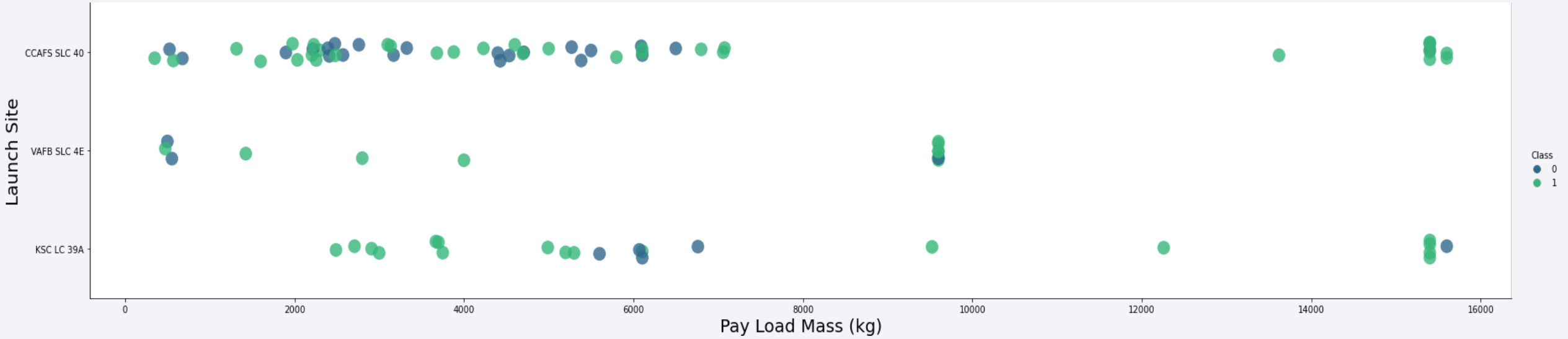
Green Dots: Successful Launch

Purple Dots: Unsuccessful Launch

Graph shows that success rate grows over time which is indicated in Flight Number.

CCAAFS seems to be the main launch site as it has the most volume.

Payload vs. Launch Site



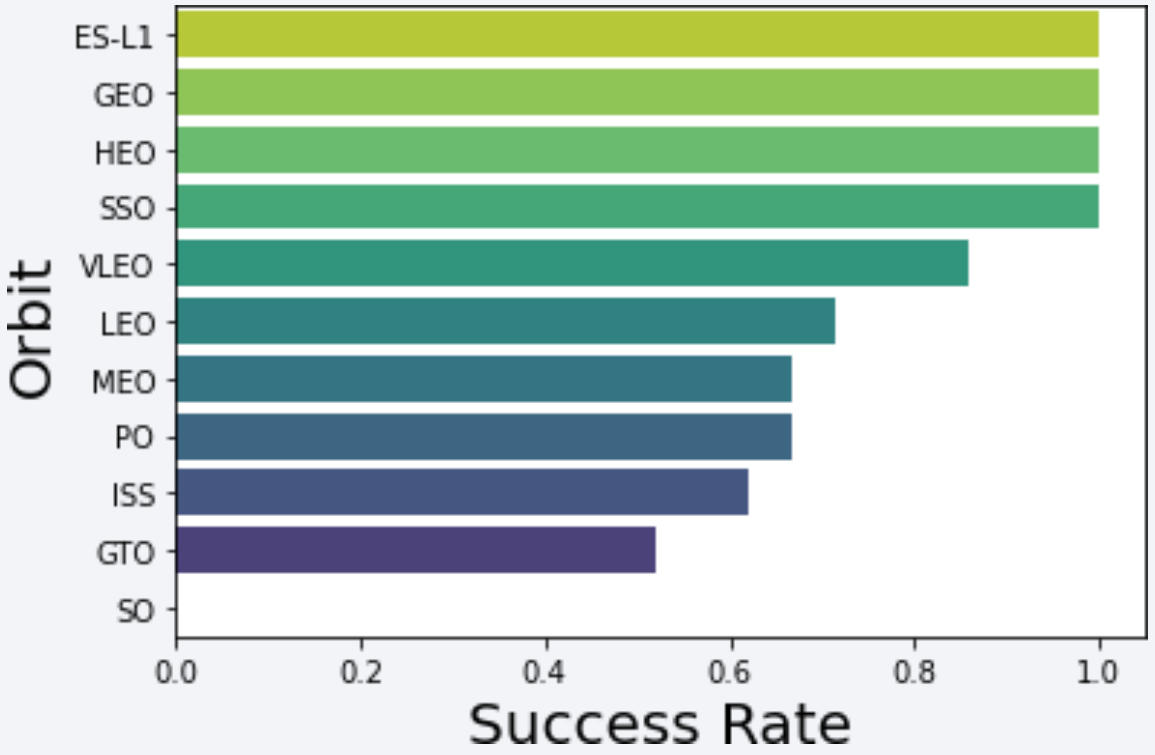
Green Dots: Successful Launch

Purple Dots: Unsuccessful Launch

Payload 0-6000 kg seems likely to fall.

It appears that different launch sites use different payload mass as well.

Success Rate vs. Orbit Type



Success Rate Scale with
0 as 0%
0.6 as 60%
1 as 100%

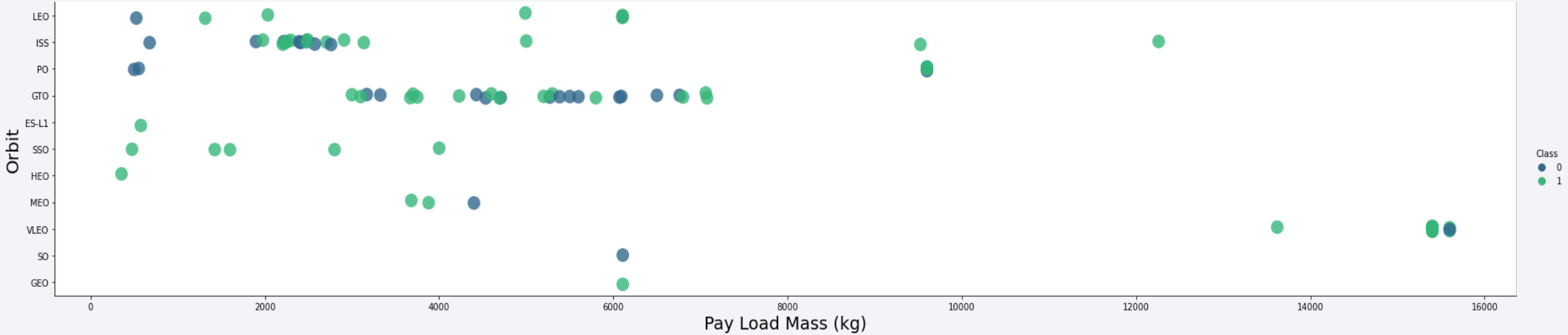
ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate

VLEO (14) has decent success rate and attempts

SO (1) has 0% success rate

GTO (27) has the around 50% success rate but largest sample

Payload vs. Orbit Type



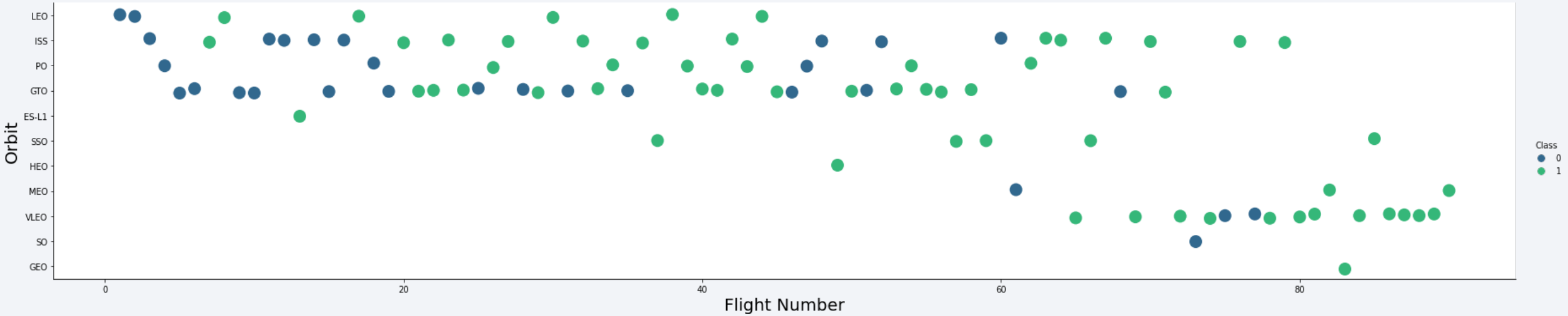
Green Dots: Successful Launch

Purple Dots: Unsuccessful Launch

Payload mass has correlation with orbit.

LEO and SSO have low Payload Mass relatively.

Flight Number vs. Orbit Type



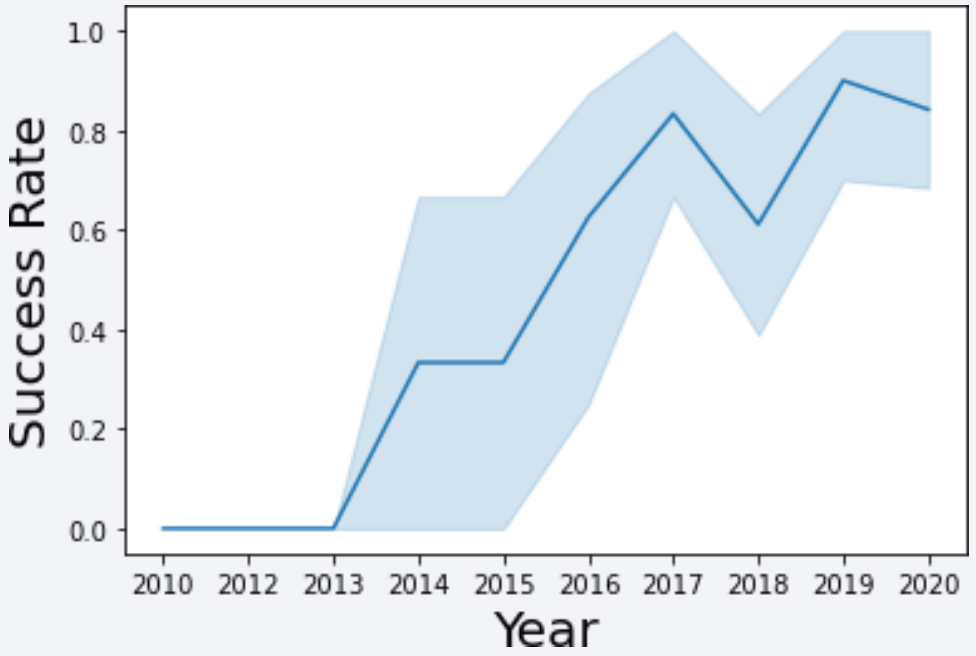
Green Dots: Successful Launch

Purple Dots: Unsuccessful Launch

Launch Orbit preferences changed over Flight Number.

Launch Outcome has correlation with this preference.

Launch Success Yearly Trend



95% confidence interval
(light blue shading)

Generally Success Rate grows yet some local minima occurred around 2018

EDA with SQL

EXPLOTARY DATA ANALYSIS with SQL DB2

INTEGRATED IN PYTHON with SQLALCHEMY

All Launch Site Names

```
%sql select DISTINCT LAUNCH_SITE FROM DB.SPACEX
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-98
Done.
```

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

Launch Site Names Begin with 'CCA'

```
%sql select * from DB.SPACEX where launch_site like 'CCA%' limit 5
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqeb8u7740.databases.appdomain.cloud:30089/bludb
Done.
```

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

Total Payload Mass

```
%sql select sum(payload_mass__kg_) as sum from DB.SPACEX where customer like 'NASA (CRS)'
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqeb8u7740.datal
Done.
```

SUM
45596

This query sums the total payload mass in kg where NASA was the customer.

Average Payload Mass by F9 v1.1

```
%sql select avg(payload_mass__kg_) as Average from DB.SPACEX where booster_version like 'F9 v1.1%'
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqeb8u7740.databases.appdo
Done.
```

average
2534

This query calculates the average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

```
%sql select min(date) as Date from DB.SPACEX where mission_outcome like 'Success'
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqeb8u7740.dat
Done.
```

DATE
2010-06-04

This query provides the dates of the first successful landing outcome

Successful Drone Ship Landing with Payload between 4000 and 6000

```
%sql select booster_version from DB.SPACEX where mission_outcome like 'Success' AND payload_mass__kg_>=4000 AND payload_mass__kg_<6000 AND
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqeb8u7740.databases.appdomain.cloud:30089/bludb
Done.
```

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

This query provides the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

```
%sql SELECT mission_outcome, count(*) as Count FROM DB.SPACEX GROUP by mission_outcome ORDER BY mission_outcome
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqeb8u7740.databases.appdomain.cloud:3
Done.
```

]:

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

This query gives count of mission outcome.

Boosters Carried Maximum Payload

```
maxm = %sql select max(payload_mass__kg_) from DB.SPACEX
maxv = maxm[0][0]
%sql select booster_version from DB.SPACEX
where payload_mass__kg_=(select max(payload_mass__kg_) from DB.SPACEX)
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqe
Done.
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqe
Done.
```

]:

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

```
%sql select MONTHNAME(DATE) as Month, landing__outcome, booster_version, launch_site
from DB.SPACEX
where DATE like '2015%' AND landing__outcome like 'Failure (drone ship)'
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c11fb794e48e.brt9d04f0cmqeb8u7
Done.
```

9]:

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

This query produces the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

There were two such occurrences

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

```
%sql select landing__outcome, count(*) as count
from DB.SPACEX
where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing__outcome ORDER BY count Desc
```

```
 * ibm_db_sa://ac2a034c:***@cfeb76f2-853d-4ee9-985b-c
Done.
```

0]:

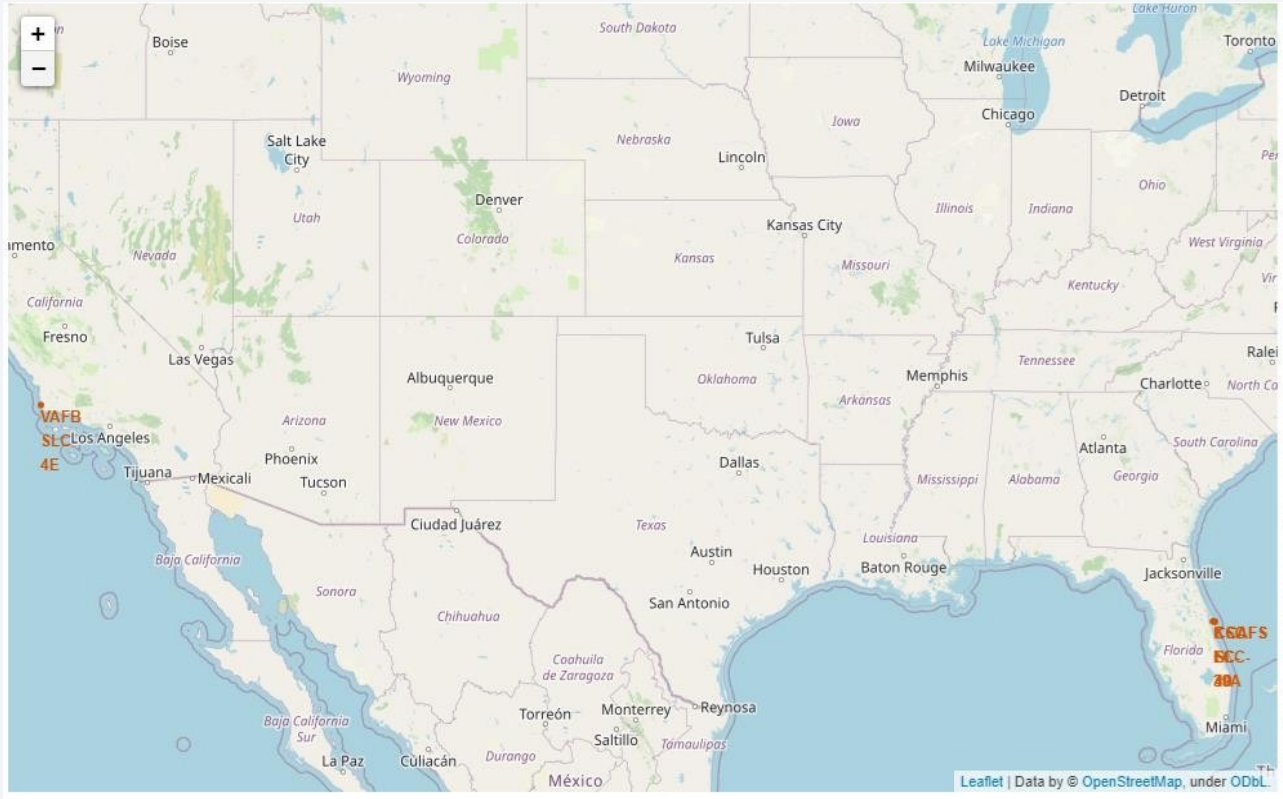
landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

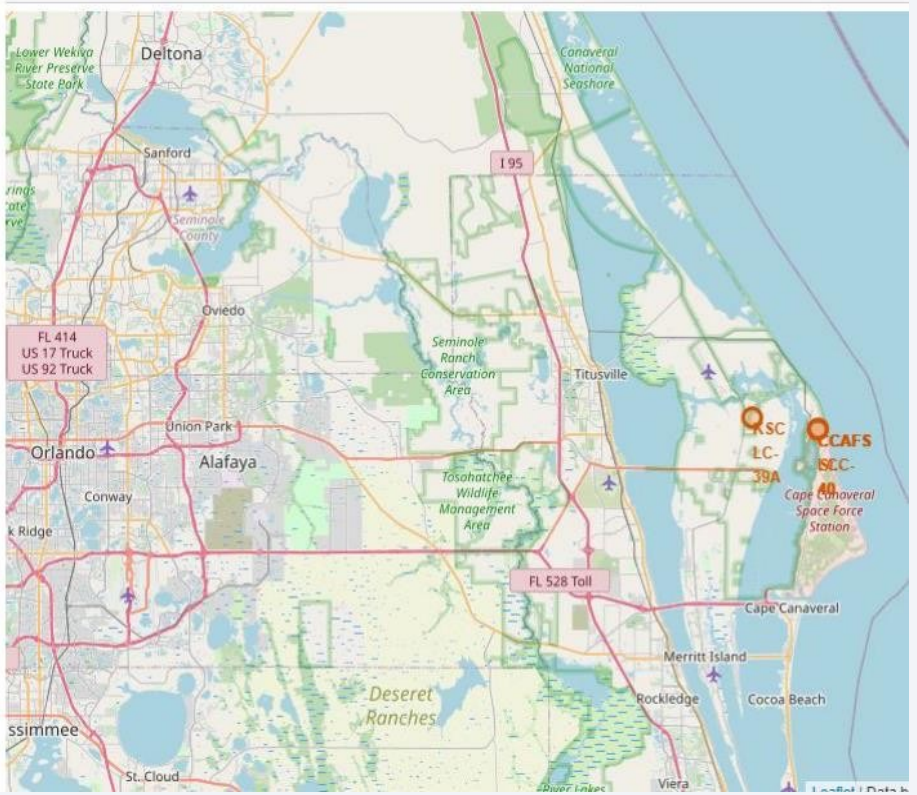
This query returns the count of landing between the date 2010-06-04 and 2017-03-20, in descending order.

Interactive Map with Folium

INTERACTIVE VISUAL ANALYTICS with Folium

Launch Site Locations



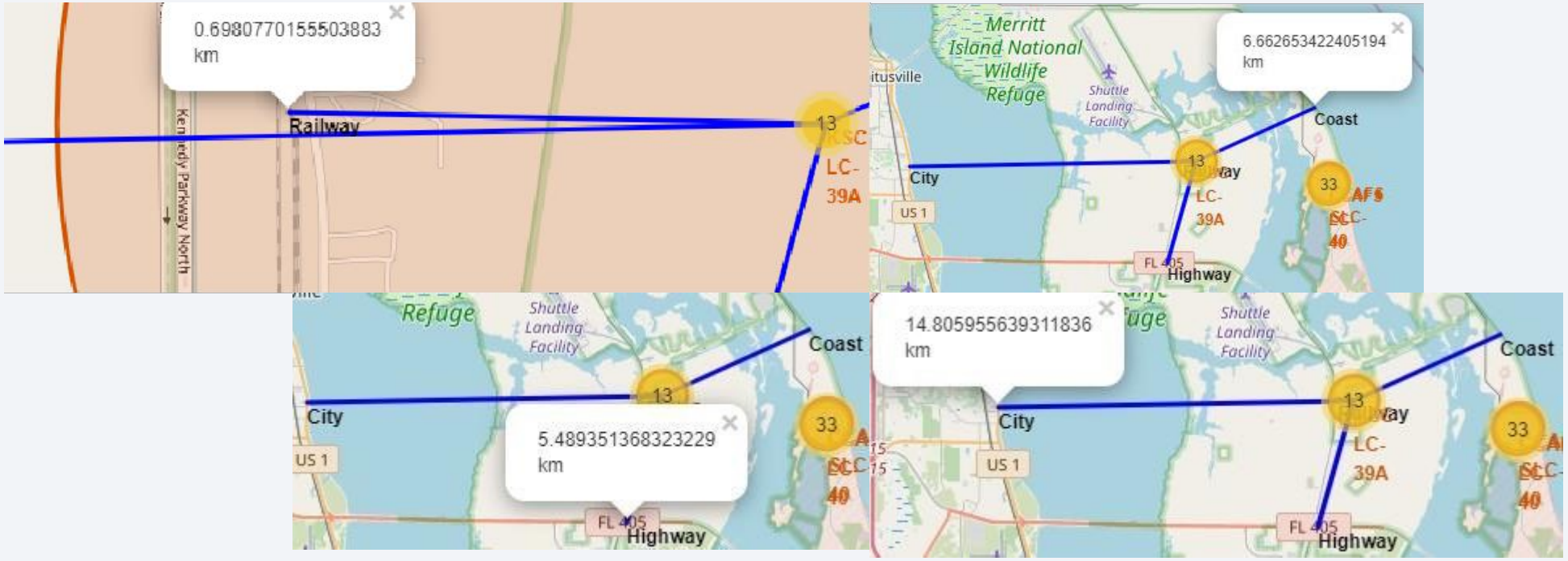


The left map demonstrates all launch sites relative US map. The right map displays the two Florida launch sites since they are very close to each other.

Color-Coded Launch Markers

On Folium Map, Clusters can be taped into to indicate each successful landing (green icon) and failed landing (red icon). In this example, VAFB SLC-4E indicates 4 successful landings and 6 failed landings.

Key Location Proximities

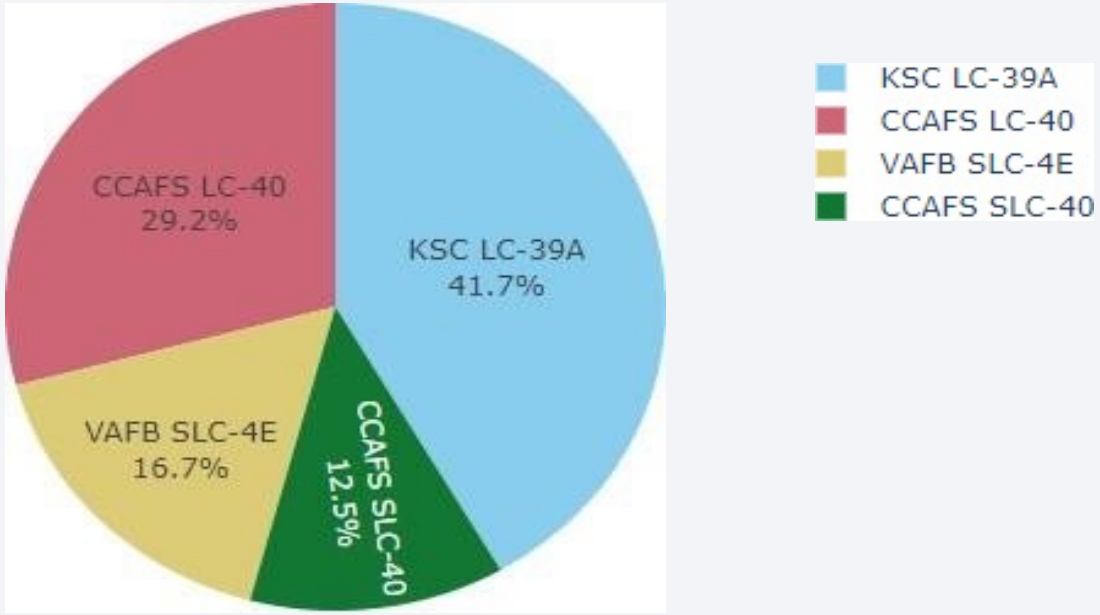


In this example KSC LC-39A, For the most part, launch sites are adjacent to trains and supply transports. Human and supply transportation are easily accessible from launch sites. Launch sites are also close to coasts and relatively far from towns, allowing launch failures to land in the sea rather than in densely populated areas.

Build a Dashboard with Plotly Dash

INTERACTIVE VISUAL ANALYTICS with Ploty Dash

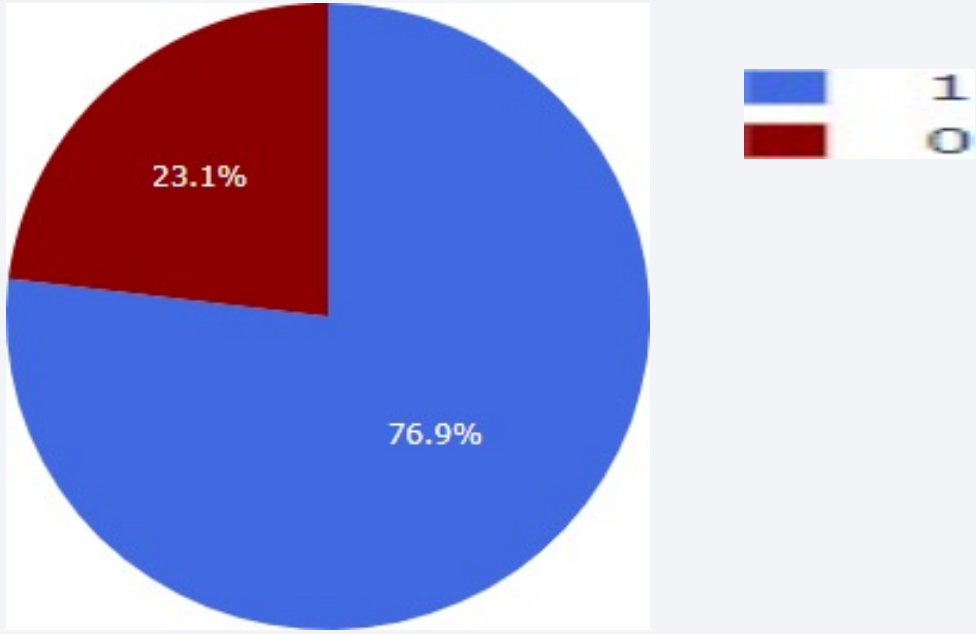
Successful Launches Across Launch Sites



This is the percentage of landings that were successful across all launch sites. CCAFS LC-40 is the old name for CCAFS SLC-40, hence CCAFS and KSC both have the same number of successful landings, although the majority of them happened before the name change. The number of successful landings at VAFB is the fewest. This could be attributed to a smaller sample size and increased launching difficulty on the west coast.

Highest Success Rate Launch Site

KSC LC-39A Success Rate (blue=success)



With 10 successful landings and 3 failed landings, KSC LC-39A has the best success rate.

Payload Mass vs. Success vs. Booster Version Category

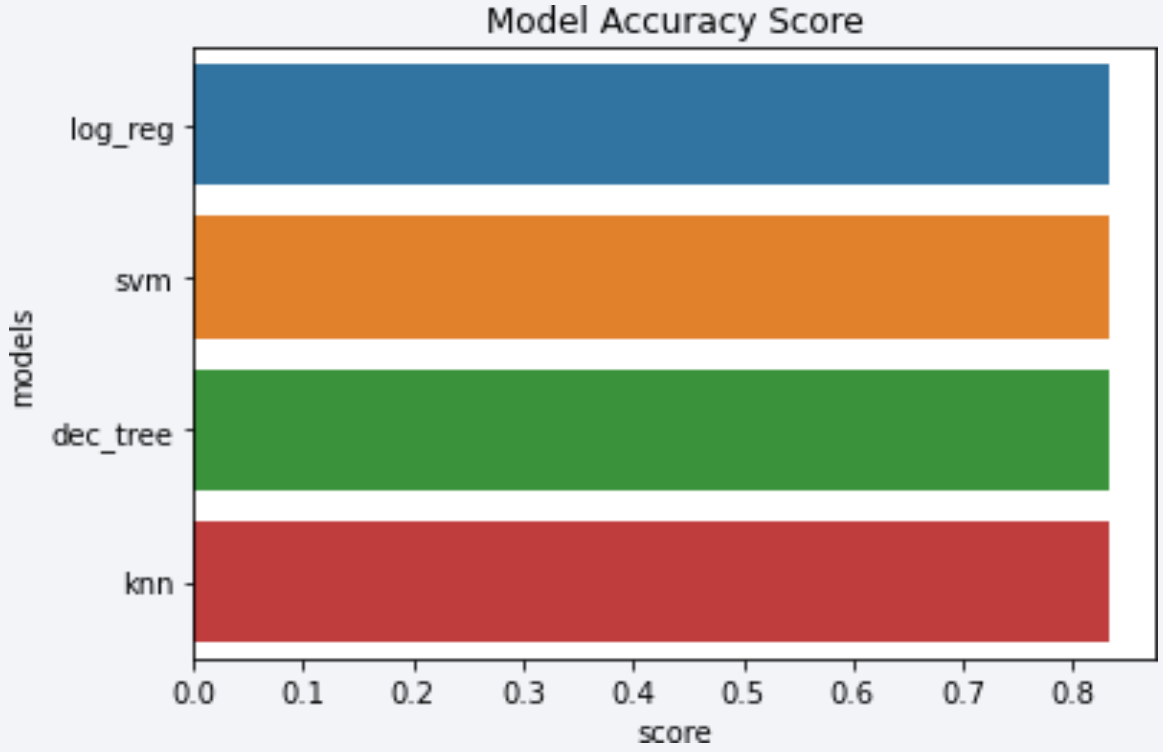
A Payload range selection is available on the Plotly dashboard. However, instead of the maximum payload of 15600, this is set to a range of 0-10000. A successful landing is indicated by a 1 and a failure by a 0. The booster version category in color and the number of launches in point size are also taken into consideration in the scatter plot. There are two unsuccessful landings with weights of zero kilograms in this exact range of 0-6000.

Predictive Analysis (Classification)

GRIDSEARCHCV(CV=10)

ON LOGISTIC REGRESSION, SVM, DECISION TREE AND KNN

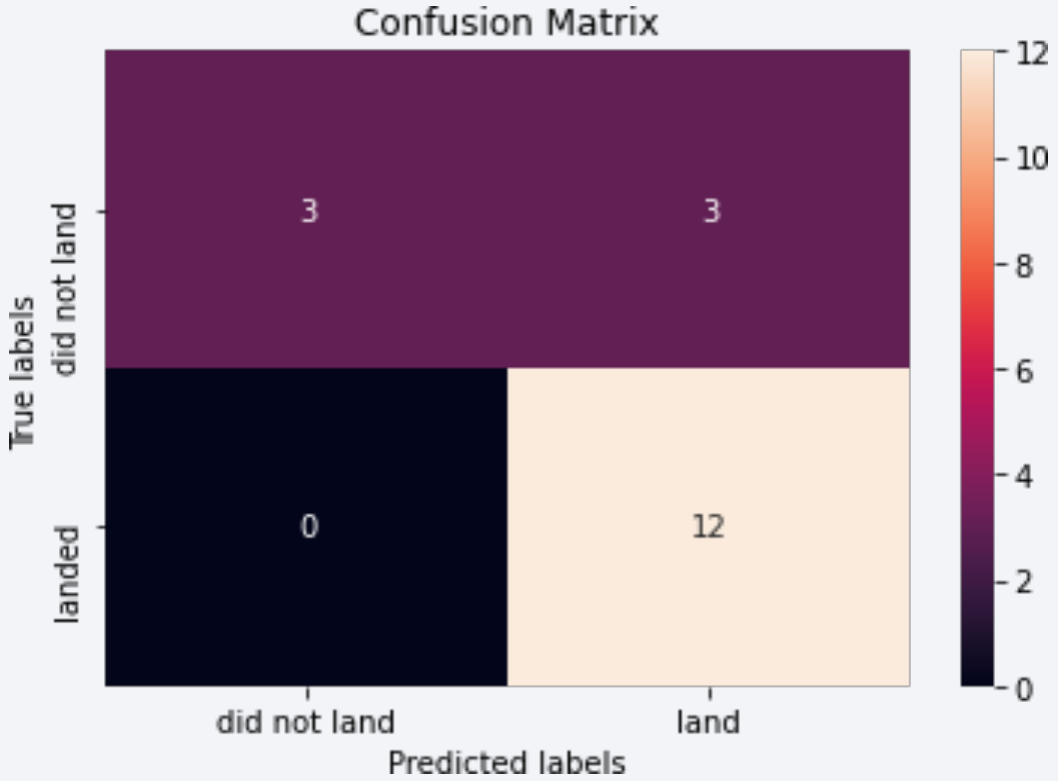
Classification Accuracy



All models have the same accuracy on the test set at 83.33% accuracy.

The sample size of data on test is %18 which is not adequate to distinguish model accuracy.

Confusion Matrix



Correct predictions are on a diagonal from top left to bottom right.

Because all models scored equally well on the test set, the confusion matrix is the same for all models.

When the true label was successful landing, the models projected 12 successful landings.

When the true label was failure landing, the models projected three unsuccessful landings.

When the true label was failed landings, the models projected three successful landings (false positives).

Conclusions

- In this project, our goal is to create a machine learning model for Space Y, which aims to compete with SpaceX in order to predict when Stage 1 will land successfully to reduce cost approximately $100 million US Dollars.
- Data is gained by a public SpaceX API and web scraping SpaceX Wikipedia page.
- Data label is generated and stored into DB2 SQL DB.
- Interactive Visual Analytical Dashboards are produced via Folium and Plotly Dash.
- Logistic Regression, Support Vector Machine, Decision Tree and KNN Machine Learning model with same an accuracy of 83%.
- The Stakeholders of Space Y company can use to predict whether wStage1 will land successfully or not.
- More data should be collected if at all feasible in order to better established machine learning model and increase accuracy.

Appendix

Github Repository URL:

https://github.com/mustafaturkoz/IBM_Data_Science_Professional_Certificate

Instructors:

Rav Ahuja, Alex Aklson, Aije Egwaikhide, Svetlana Levitan, Romeo Kienzler, Polong Lin, Joseph Santarcangelo, Azim Hirjani, Hima Vasudevan, Saishruthi Swaminathan, Saeed Aghabozorgi, Yan Luo

Special Thanks:

Instructors, IBM and Coursera