Space X

IBM Data Science Capstone Project

Majd Abu Khalaf

27/10/2021

IBM Developer

SKILLS NETWORK

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

EXECUTIVE SUMMARY

- Using Python, we have analyzed data set we get from SpaceX and Wikipedia , to build machine learning model that will help us in prediction the success of falcon 9 stage 1 landing .
- From the data we found that space x has 4 lunch sites in USA near the coastline in the west and east , and far from cities
- The landing success rate improved over years, and it reaches now more than 85% .
- We found that lower payloads lunches has higher landing success rate .
- Lunches to ES-L1 ,GEO ,HEO,SSO orbits has the best success rate .
- Finally, we reached to algorithm estimate the success of stage 1 landing with around 90% of accuracy .

INTRODUCTION

- 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.
- we will predict if the Falcon 9 first stage will land successfully.
- This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

METHODOLOGY

We will see the following methodologies we used :

- Data collection methodology
- data wrangling methodology
- EDA and interactive visual analytics methodology
- Predictive analysis methodology

Data collection methodology

- Data collection was done from two sources :
  - Space X REST API ,
  - Wikipedia “List of Falcon 9 and Falcon Heavy launches”
- We performed JSON normalization and data wrangling for the data we get through the REST API , and finally the data was imported into pandas data frame.
- We also used web scarping (beautiful soap) to arrange the HTML tables got from Wikipedia into pandas data frame .

Data wrangling methodology

- In this part we tried to review the data we have so to find out some insights for this data to help us on the next steps .
- We found the following :
  - the number of launches on each site .
  - the number and occurrence of each orbit .

```
# 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
```

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

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
HEO       1
SO        1
ES-L1     1
GEO       1
Name: Orbit, dtype: int64
```

Data wrangling methodology

- the landing outcomes .

```
landing_outcomes = df['Outcome'].value_counts()

landing_outcomes
```

```
True ASDS      41
None None      19
True RTLS      14
False ASDS      6
True Ocean      5
None ASDS       2
False Ocean     2
False RTLS      1
Name: Outcome, dtype: int64
```

- Finally we created new column we call it “Class” , that indicates if it was successful or failed landing , with Class=0 represents failed landing and Class=1 represents successful landing .

EDA and interactive visual analytics methodology

- In this part of our analysis we have done the following :
  - Import the CSV file from the pervious step to SQL server .
  - Using magic Sql we have performed analysis to find more insights form our data set like :
    - The names of lunch site.
    - Total payload mass carried for a certain customer
    - Average payload carried by falcon 9 booster version 1.1
    - The first successful landing date
    - And more

EDA and interactive visual analytics methodology

- Also we used seaborn to Visualize relations between different inputs in the dataset we have on hand like :
  - relationship between Flight Number and Payload Mass
  - relationship between Flight Number and Launch Site
  - relationship between Payload and Launch Site
  - relationship between success rate of each orbit type
  - relationship between Flight Number and Orbit type
  - relationship between Payload and Orbit type
  - Visualize the launch success yearly trend

- We used the function get_dummies on the dataframe to apply OneHotEncoder to the column Orbits, LaunchSite, LandingPad, and Serial.

EDA and interactive visual analytics methodology

- We also preformed more interactive visual analytics using Folium.
- Using folium the following was done :
  - Mark all launch sites on a map
  - Mark the success/failed launches for each site on the map
  - Calculate the distances between a launch site to its proximities

- To have more insights about the relation between different inputs we have created dashboard using Plotly dash .
- In the dashboard we viewed the following :
  - Success rate for each lunch site
  - Success for each payload and booster version in the selected lunch site .

Predictive analysis methodology

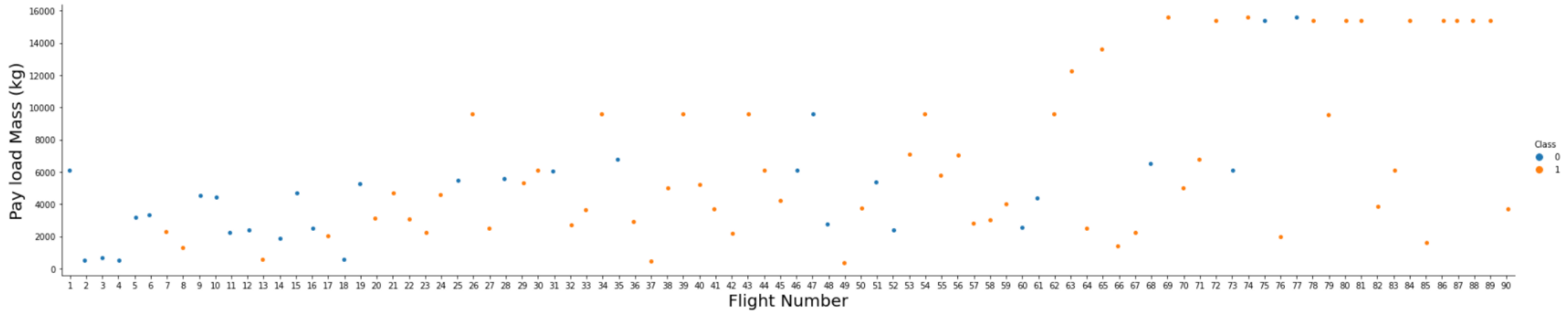
- In order to finally build our prediction model we have done the following :
  - Load the data frame and name it X
  - Create a NumPy array from the column Class in data frame and assign it to Y
  - Standardize the data in X then reassign it to the variable X using the transform
  - Use the function train_test_split to split the data X and Y into training and test data.
  - Test 4 models :(logistic regression , support vector machine, tree classifier,k nearest neighbours)
  - and find our the best parameters for each model using GridSearch .
  - and the accuracy of each model using method score .
  - Plot the confusion matrix for each model prediction .

RESULTS

We will see the following Results we found :

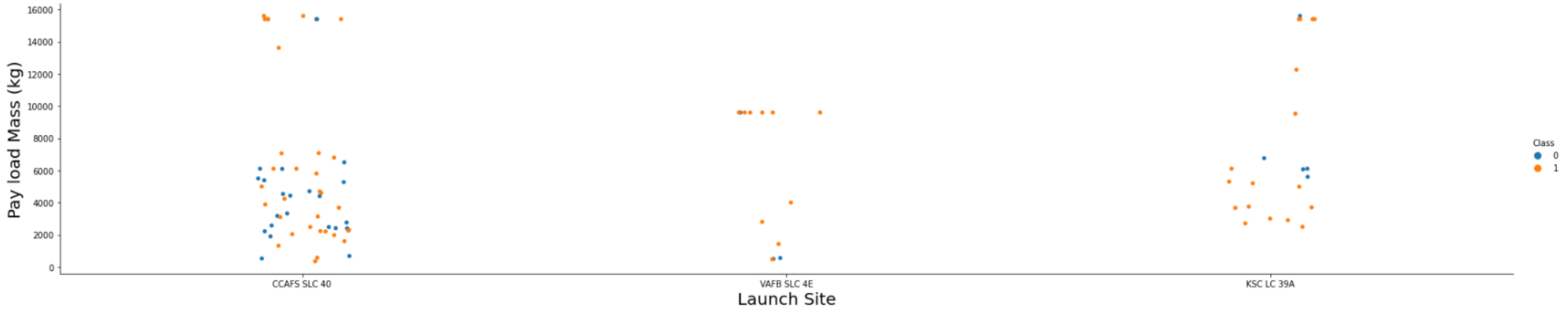
- EDA with visualization results
- EDA with SQL results
- interactive map with Folium results
- Plotly Dash dashboard results
- Predictive analysis (classification) results

EDA with visualization results



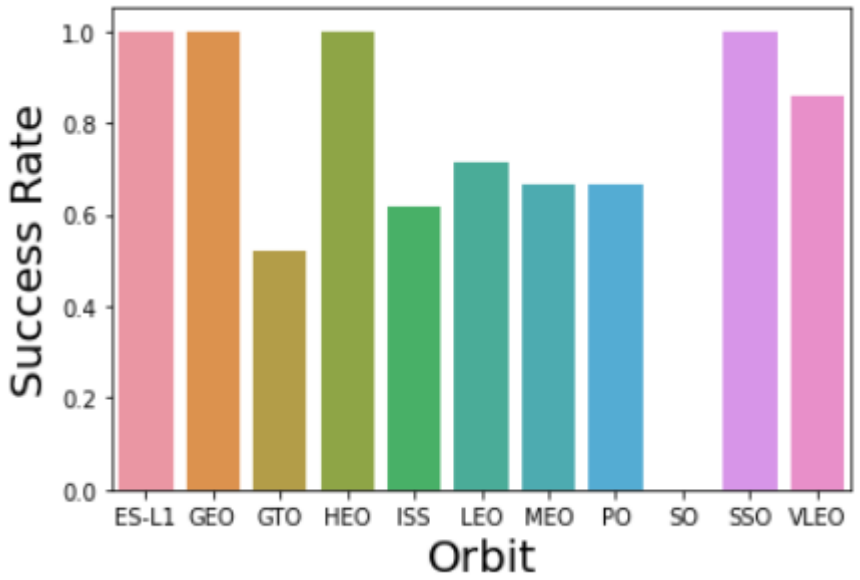
- Plotting out the Flight Number vs. Payload Mass and overlay the outcome of the launch.
- We see that:
  - as the flight number increases, the first stage is more likely to land successfully.
  - The payload mass is also important; it seems the more massive the payload, the less likely the first stage will return.

EDA with visualization results



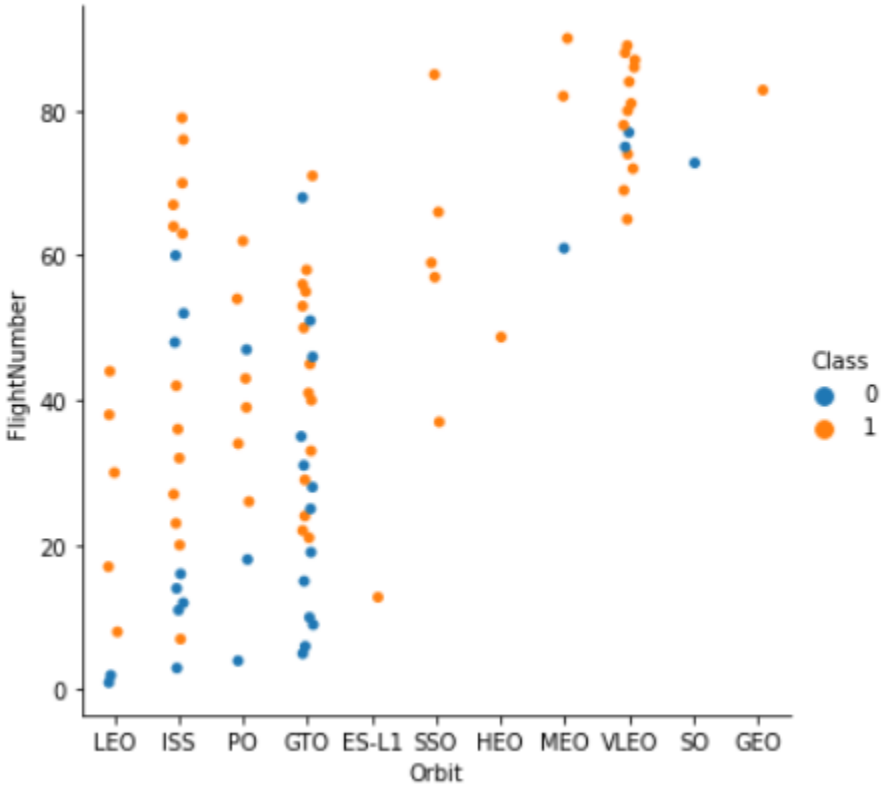
- Plotting out the Lunch site vs. Payload Mass and overlay the outcome of the launch.
- We see that :
  - CCAFS SLC 40 handles the highest payload lunches and the landing success rate is higher on the bigger payloads lunched from this site .
  - VAFB SLC 4E has the lowest number for lunches with the highest landing success rate
  - KSC LC 39A ranked 2nd in the number of lunches and 2nd in terms of success rate , it also handles big payload , but lower than CCAFS SLC 40 in terms of number of lunches , also it shows high success rate for the payloads between 3000 to 6000 .

EDA with visualization results



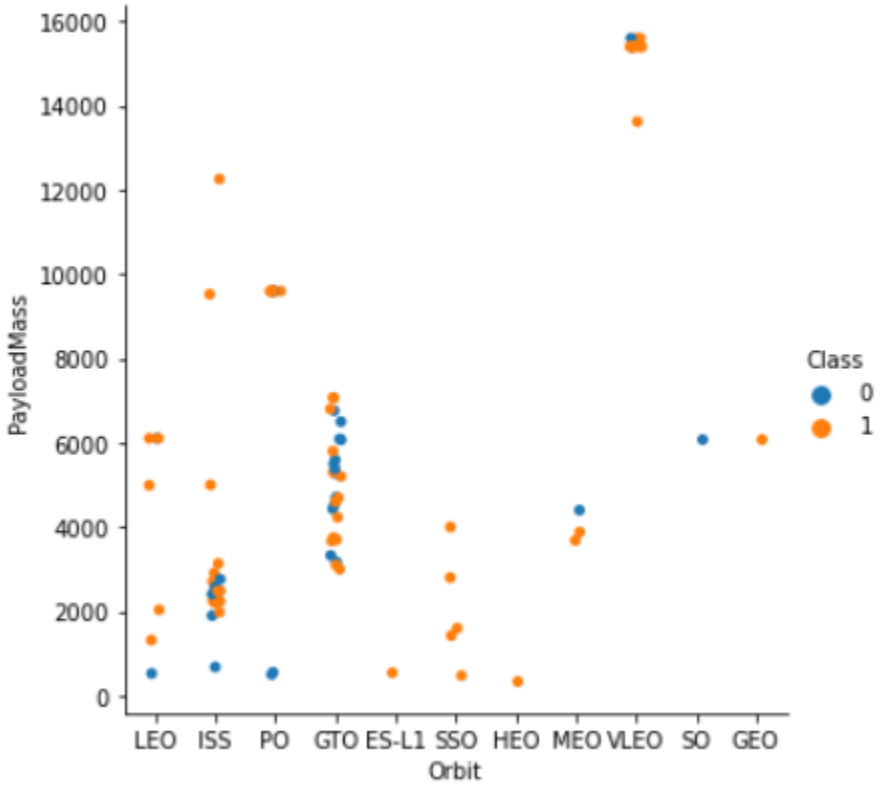
- Comparing the landing success rate with the target lunch orbit we found that (ES-L1 ,GEO ,HEO,SSO) orbits has the highest success rate 100% .
- GTO orbit shows the low success rate ~ 52% .
- SO orbit has no successful landing for the Stage 1.

EDA with visualization results



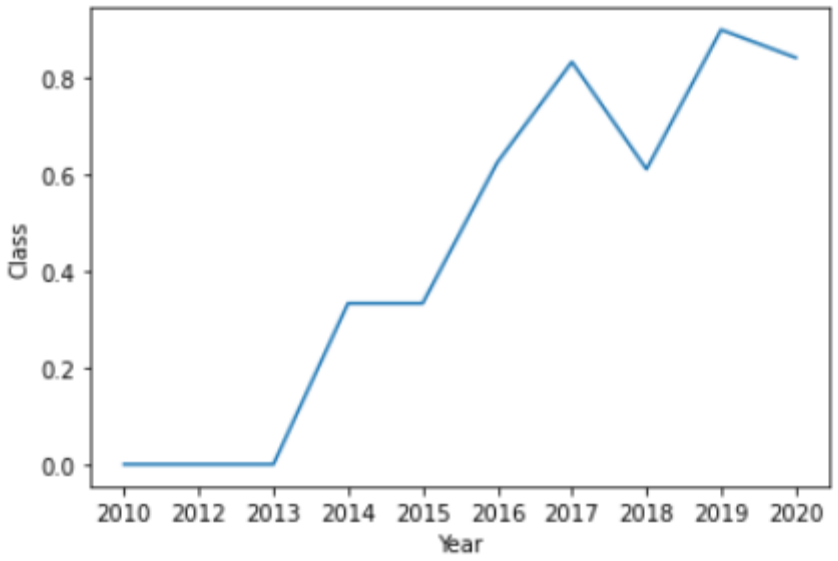
- We see that in the LEO orbit the Success appears related to the number of flights.
- on the other hand, there seems to be no relationship between flight number when in GTO and ISS orbits.
- Lunching to VLEO orbit was in later flight numbers .

EDA with visualization results



- We observe that Heavy payloads have a negative influence on GTO orbits.
- Heavy payloads have positive on LEO and ISS orbits.
- The most heavy payloads was lunched to VLEO Orbit

EDA with visualization results



- We can observe that the success rate since 2013 started improving
- we see in the highest success rate reached in 2019 .

EDA with SQL results

- The names of the unique launch sites in the space mission

```
%sql SELECT DISTINCT launch_site FROM SPACEXDATASET
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-4
Done.
```

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

- The total payload mass carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM(payload_mass__kg_) FROM SPACEXDATASET WHERE customer='NASA (CRS)'
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90108kqb1od81
Done.
```

1
45596

EDA with SQL results

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

```
%sql SELECT AVG(payload_mass__kg_) FROM SPACEXDATASET WHERE booster_version='F9 v1.1'
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90108kqb1od8lcg.
Done.
```

1
2928

- The date when the first successful landing outcome in ground pad was achieved

```
%sql SELECT min(DATE) FROM SPACEXDATASET WHERE landing__outcome='Success (ground pad)'
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90108kqb1od8lcg.
Done.
```

1
2015-12-22

EDA with SQL results

- 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 SPACEXDATASET  WHERE landing__outcome='Success (drone ship)' and payload_mass__kg_ >4000 and pa
yload_mass__kg_ < 6000
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

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

- The total number of successful and failure mission outcomes

```
%sql select count(landing__outcome) from SPACEXDATASET where landing__outcome like 'Success%' or landing__outcome like 'Failur
e%'
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

1
71

EDA with SQL results

- The names of the booster_versions which have carried the maximum payload mass

```
%sql select DISTINCT booster_version , payload_mass__kg_ from SPACEXDATASET WHERE payload_mass__kg_ = (select max(payload_mass__
kg_) from SPACEXDATASET)
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

booster_version	payload_mass__kg_
F9 B5 B1048.4	15600
F9 B5 B1048.5	15600
F9 B5 B1049.4	15600
F9 B5 B1049.5	15600
F9 B5 B1049.7	15600
F9 B5 B1051.3	15600
F9 B5 B1051.4	15600
F9 B5 B1051.6	15600
F9 B5 B1056.4	15600
F9 B5 B1058.3	15600
F9 B5 B1060.2	15600
F9 B5 B1060.3	15600

EDA with SQL results

- the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select landing__outcome,booster_version , DATE from SPACEXDATASET where landing__outcome like 'Fail%' and YEAR(DATE)=2015
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

landing__outcome	booster_version	DATE
Failure (drone ship)	F9 v1.1 B1012	2015-01-10
Failure (drone ship)	F9 v1.1 B1015	2015-04-14

- the count of landing outcomes between the date 2010-06-04 and 2017-03-20, ranked in descending order .

```
%sql select  landing__outcome,count(landing__outcome) as rank  from SPACEXDATASET where DATE > '2010-06-04' and DATE <'2017-03-20' group by landing__outcome order by rank desc
```

```
 * ibm_db_sa://ptd36129:***@764264db-9824-4b7c-82df-40d1b13897c2.bs2io90108kqb1od8lcg.databases.appdomain.cloud:32536/bludb
Done.
```

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

Interactive map with Folium results

- Using folium we draw all launch sites on a map

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

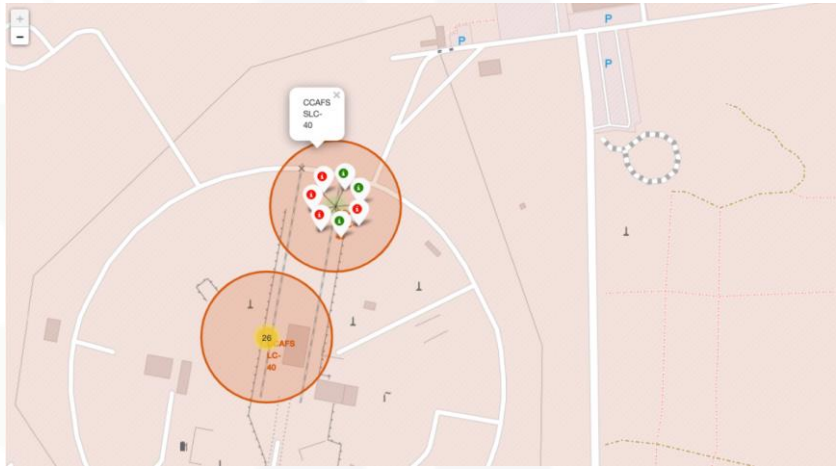
Interactive map with Folium results

- Using folium we draw all launch sites on a map

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

Interactive map with Folium results

- Using folium / MarkerCluster we Mark the success/failed launches for each site on the map



Green : for Successful landing

Red : for Failed landing

Interactive map with Folium results

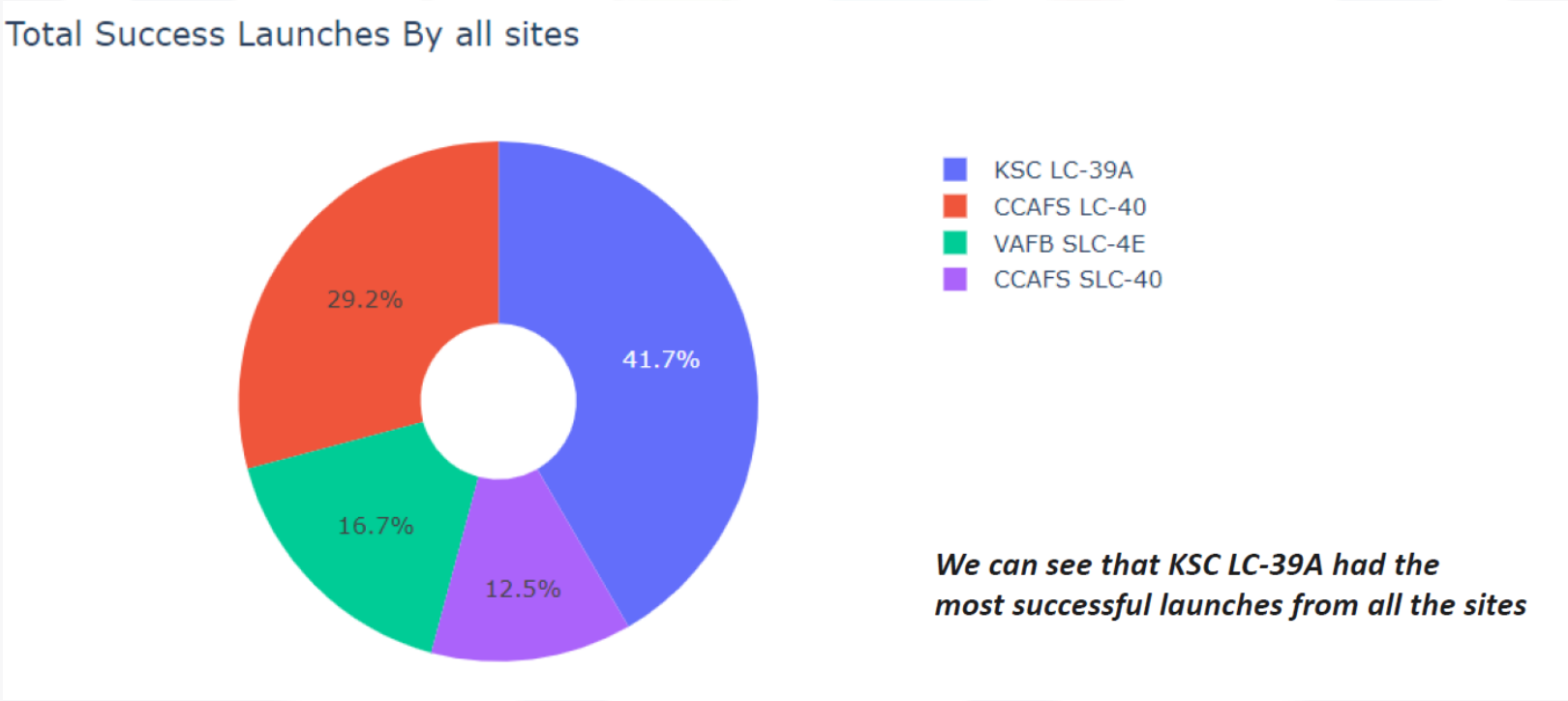
- Using folium calculate the distance between the lunch site and the sea or highways or cities :

- Are launch sites in close proximity to railways? NO
- Are launch sites in close proximity to highways? NO
- Are launch sites in close proximity to coastline? YES
- Do launch sites keep certain distance away from cities? YES

- Looking on the above answers we see that the main feature in the lunching site , it to be away from population and near the costal line , which is logical , to avoid any problems coming from such activity .

Plotly Dash dashboard results

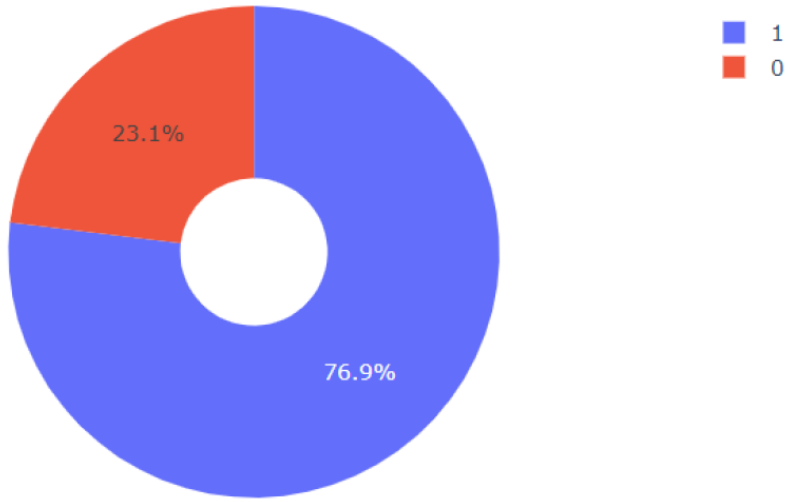
- Pie chart represents the percentage from total success for all sites



We can see that KSC LC-39A had the most successful launches from all the sites

Plotly Dash dashboard results

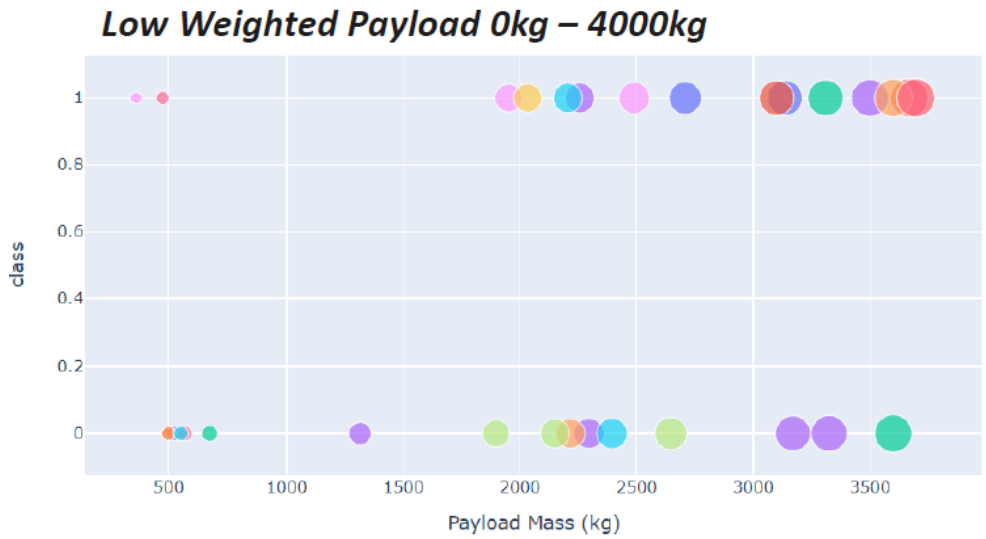
- Pie chart represents the percentage of successful and failed lunches from Site KSC LC-39A

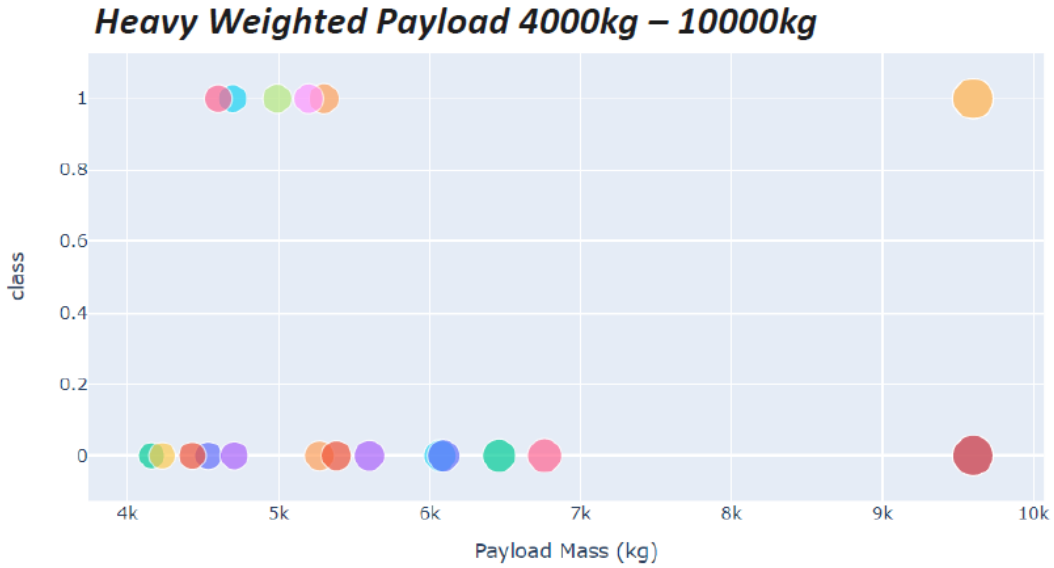


KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Plotly Dash dashboard results

- Load vs lunch outcome for different booster models





We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Predictive analysis (classification) results

- We tried the following model to estimate the if the landing is successful or not :
  - Logistic regression (logreg)
  - Support vector machine (SVM)
  - Decision tree classifier (Tree)
  - k nearest neighbours (KNN)

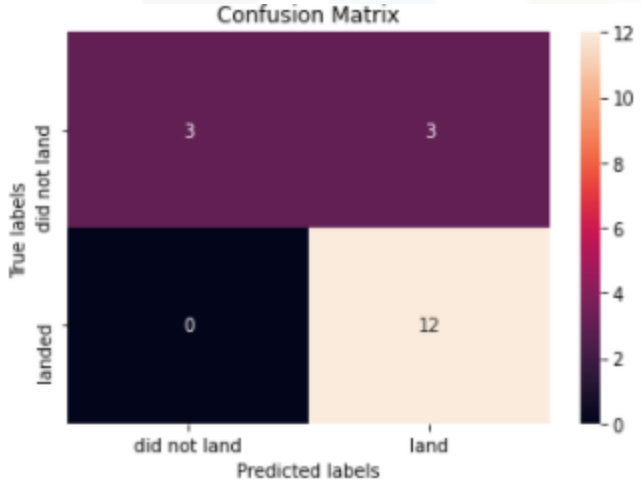
- Looking at the best score achieved from each model we found that Tree model achieved the highest accuracy .

```
algo={'logreg':logreg_cv.best_score_,'SVM':svm_cv.best_score_,'Tree':tree_cv.best_score_,'KNN': KNN_cv.best_score_}
print(algo)
max(algo, key=algo.get)
```

```
{'logreg': 0.8464285714285713, 'SVM': 0.8482142857142856, 'Tree': 0.8910714285714285, 'KNN': 0.8482142857142858}
```

```
]: 'Tree'
```

Predictive analysis (classification) results



- Looking at the confusion matrix for Decision tree classifier model , we see that the successful landing was correctly predicted .
- But the failed landing has some errors , 50% of the prediction was correct .

CONCLUSION

- Decision tree classifier algorithm is the best machine learning algorithm for the estimation of successful landing .
- Lower weight payloads lunches performs better , than the higher weight payloads in terms of successful landing for stage 1 .
- The success rate of landing of space x falcon 9 stage 1 increased over time starting from 2013 .
- VAFB SLC 4E lunch site has the best Success Rate .
- Orbit GEO,HEO,SSO,ES L1 has the best Success Rate