Data Science with SPACEY

Deep Doshi

17 March 2023

OUTLINE

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

EXECUTIVE SUMMARY

- The following methods were used in the analysis of the SpaceX Launch Data
  - Data collection using the SpaceX API and web-scrapping.
  - EDA, data wrangling, visualizations analysis and interactive visualization analytics were performed on the data.
  - Machine Learning models were built for making predictions for the success of a launch.

- Results
  - Analysis of the data revealed that some features which were more useful in making predictions for a launch.
  - Multiple models were built and the best one was selected for the predictions.

INTRODUCTION

- The objective is to evaluate the cost and the factors that would affect the success of the launches for a new rival company SpaceY.

- Required Results
  - The location for the launch sites.
  - More information on the surroundings of these sites.
  - An efficient way to predict the success or failure of a launch and its cost.

METHODOLOGY

- Data Collection:
  - Data was collection using 2 methods.
    - SpaceX API (https://api.spacexdata.com/v4/rockets/)
    - Web Scraping (https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches).

- Data Wrangling.
  - Data was modified to have a result column denoting a launch's success (1) or failure (0).

- Performed EDA on the data using SQL, Pandas, visualizations.
  - Success rates for different launch sites, orbits, payloads throughout the years.

METHODOLOGY

- Generate interactive visualizations using Folium and interactive web apps using Plotly and Dash.
  - Launch Site analysis.
  - Success rates for a selected launch site and a range of payload mass.

- Performing Predictive Analysis using Machine Learning.
  - Transformed data was normalized.
  - Divided into Test and Training sets.
  - The training set was used to build various models which were the compared based on their accuracy.

DATA COLLECTION

- Data was requested from the SpaceX API.
  - SpaceX offers a public API using which we can obtain data on their rocket launch data.
  - The data was then filtered to focus only on the **Falcon9** launches.
  - Missing values were handled by replacing them with mean and modes based on column types.

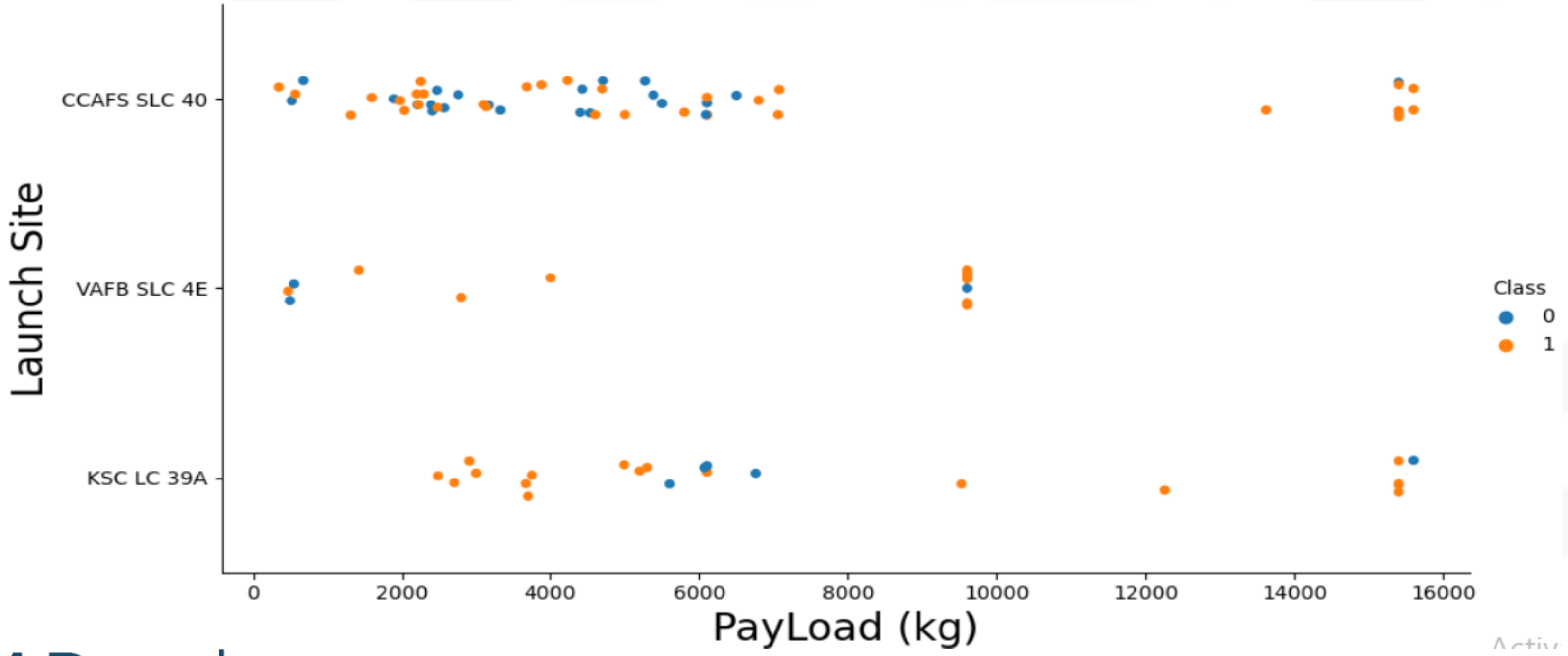
- And it was also scrapped from the Wikipedia.
  - Same information is also present on Wikipedia page.
  - The page was requested using the **BeautifulSoup** package.
  - The HTML Tables were parsed to convert the values to create the data frames.

DATA WRANGLING

- Analysis was performed on the data.
- Success rates for different launch sites, orbits, payloads and throughout the years.
- Finally, a result column was created that held the information about the success and failure of the launch.

EDA with DATA VISUALIZATION

- Exploration was performed using bar plots, scatter plots for understanding the relations between pairs of features.
  - The pairs of features were Launch Site and Payload, success rate and Orbit Types, Orbit Types and Flight Numbers etc.



EDA with SQL

- Exploration on the data was also done using SQL. Following queries were performed:
  - The names of the unique launch sites.
  - 5 launch sites that begin with the string 'CCA'.
  - Total payload mass carried by boosters launched by NASA (CRS).
  - Average payload mass carried by booster version F9 v1.1.
  - Date of the first successful landing outcome in ground pad.
  - Successful boosters in drone ship that have payload mass between 4000 and 6000.
  - Total number of successful and failure mission outcomes.
  - Names of the booster versions which have carried the maximum payload mass.
  - Failure for drone ship ,booster versions, launch site and months for the months in year 2015.
  - Rank the count of successful landing outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

INTERACTIVE MAP Using FOLIUM

- Folium maps with Circles, Markers, MarkerClusters, MousePositions were generated.
  - Circles are used to highlight areas surrounding the launch sites like Cape Canaveral Space Launch Complex 40 (CCAFS LC-40).
  - Markers are used to mark the co-ordinates of the launch sites.
  - Mouse Positions were used to calculate the co-ordinates of the location the mouse is pointing to on the map.
  - Lines were used to display the distance between the launch sites and other locations such as railways, coastlines and cities etc.

INTERACTIVE DASHBOARD Using DASH

- Graphs were displayed on an Interactive UI to visualize the data.
    - Pie chart to display the success and failure rate of a selected Launch Site.
    - Payload range slider to select the launches in the specified range of payloads to analyze.

- The dashboard allows effortless analysis of the relation between payload ranges, launch sites and their success and failure rates.

PREDICTIVE ANALYSIS Using ML

- Built ML models to train on the data for prediction of launch success or failure.
  - Decision Tree
  - Logistic Regression.
  - Support Vector Machines
  - K Nearest Neighbours.

- The data was standardized and split into training and testing sets.

- Hyper-parameter optimization was done on the models to find the best parameters for the models.

- The accuracy scores of the models were compared to select the best one.

RESULTS

- EDA Results:
  - SpaceX uses 4 different launch sites. CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, CCAFS SLC-40.
  - The average payload of F9 v1.1 booster is 2,928.4 kg.
  - The total payload of NASA Boosters is 45,596 kg.
  - The first successful ground pad landing was done on 1st May 2017.
  - Only 1 in-flight launch resulted in failure. Rest all were a success.
  - Only 2 drone ship failures were reported in the F9 v1.1 B1012 and F9 v1.1 B1015 boosters.
  - The success rate for the launches have increased over the years after the year 2013.

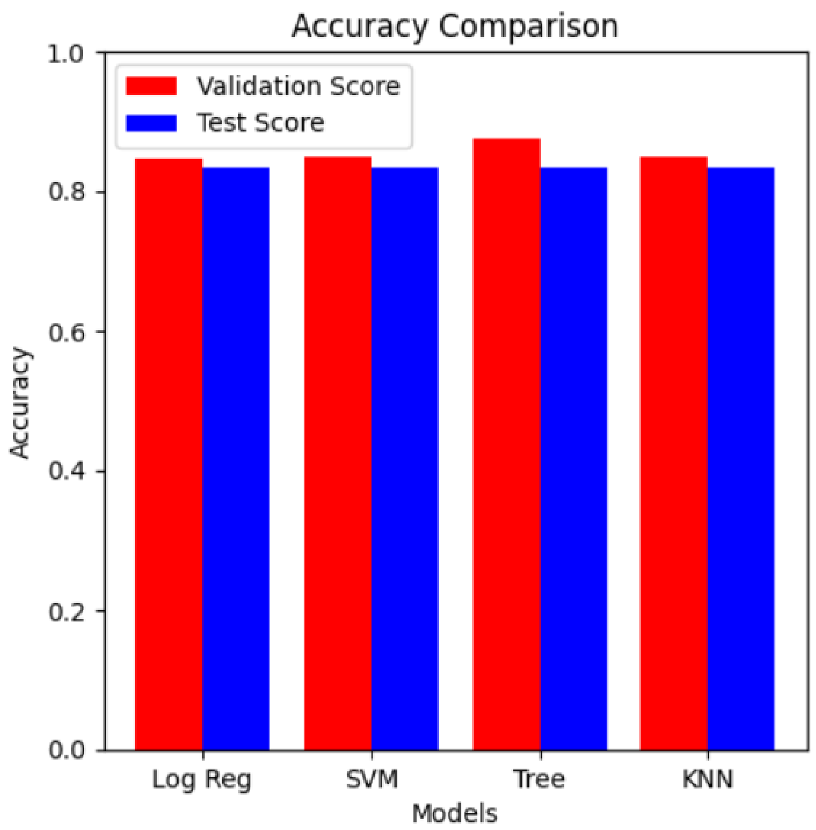
RESULTS

- Interactive folium maps showed that most of the successful launches were near the coastlines away from cities in safety locations.

- These locations also have sophisticated infrastructure such as railways.

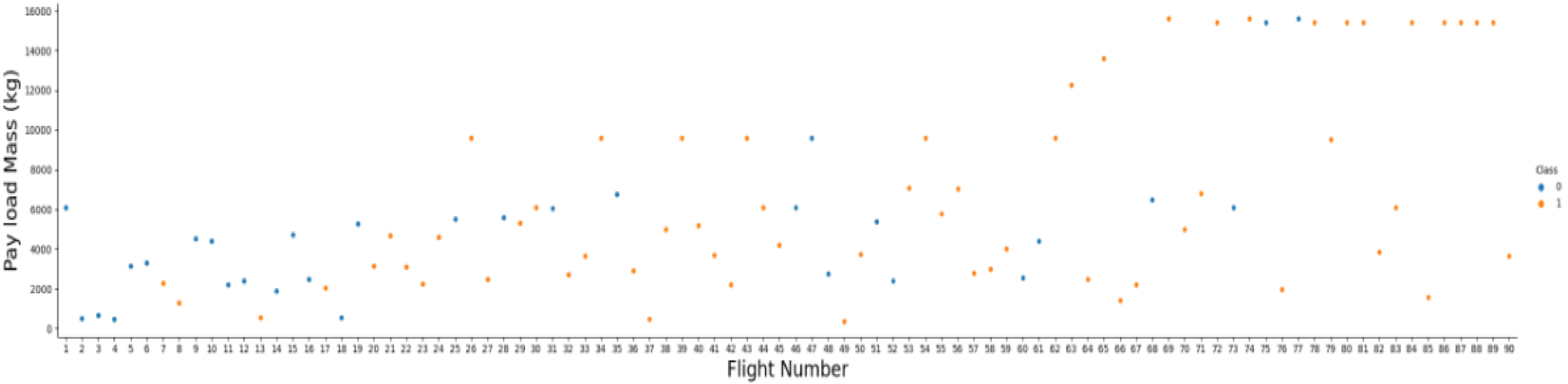
RESULTS

- The predictive analysis revealed that the Decision Tree is the best model for the predictions as it had the highest accuracy score.



PAYLOAD vs FLIGHT NUMBER

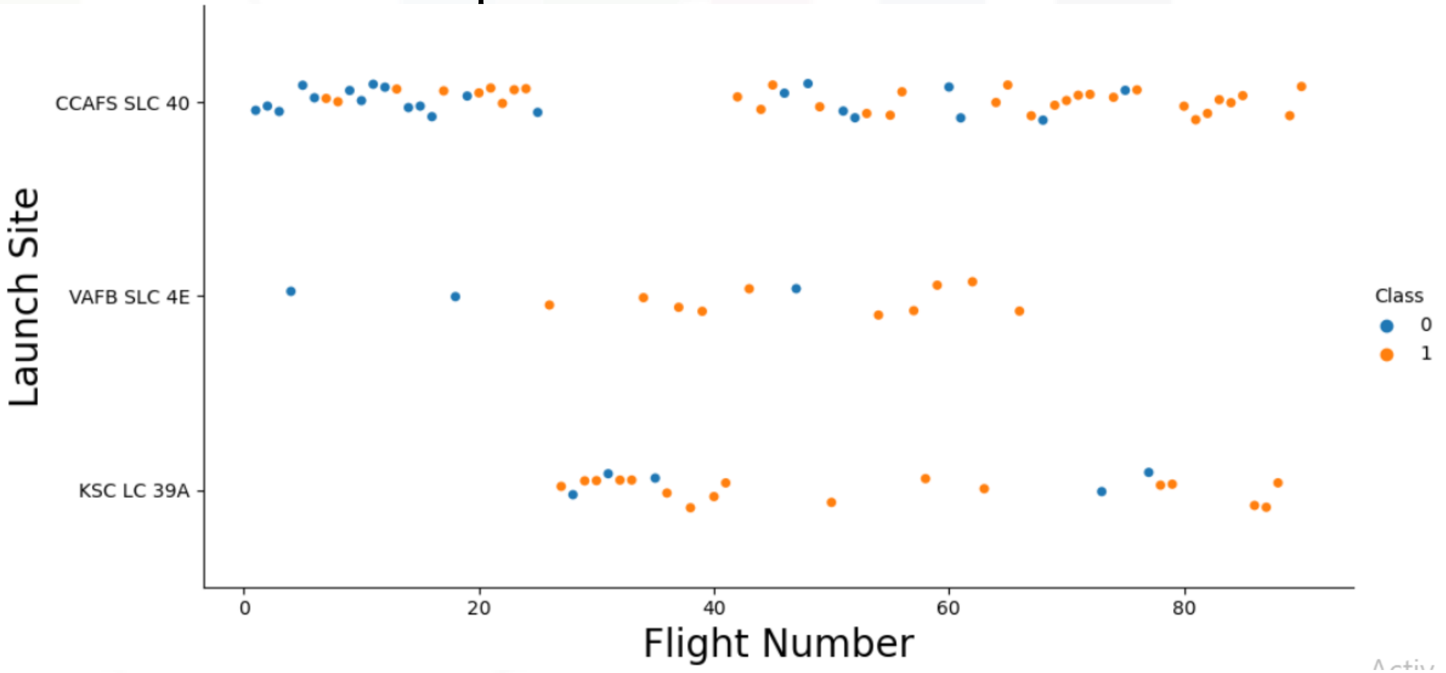
- Different launch sites have different success rates.
- CCAFS LC-40, has a success rate of 60 %.
- KSC LC-39A and VAFB SLC 4E has a success rate of 77%.



LAUNCH SITE vs FLIGHT NUMBER

- The CCAFS SLC 40 has the highest success rate.

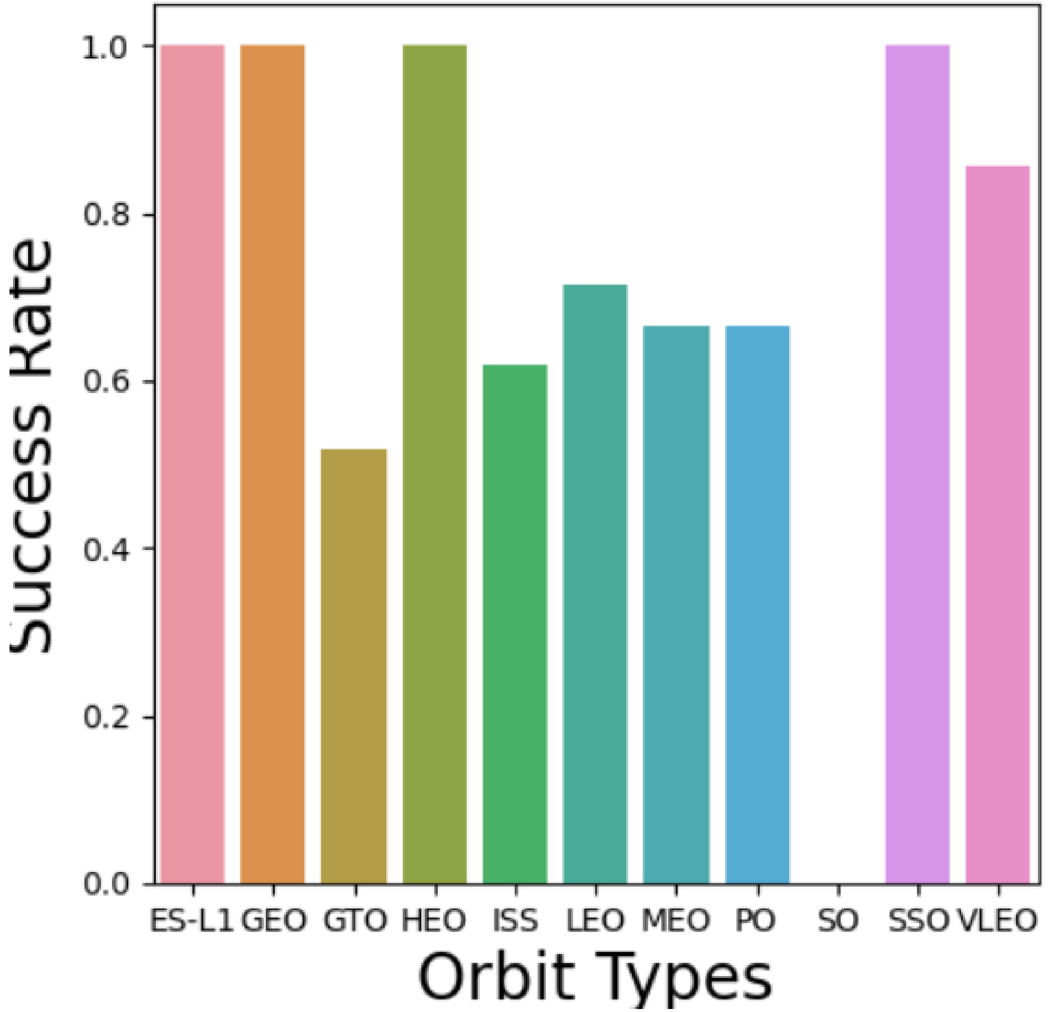
- The success rates improved with more launches performed.



SUCCESS RATE vs ORBIT TYPE

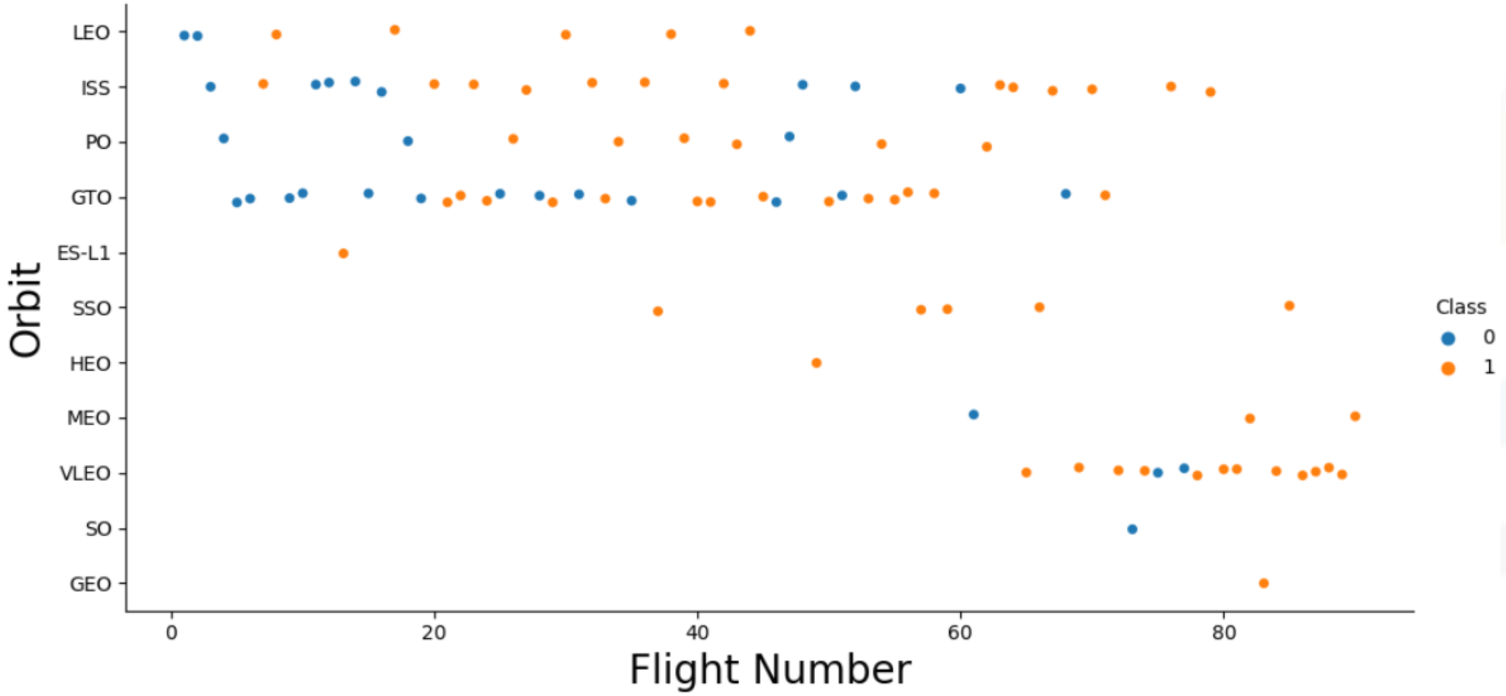
- The highest success rate is for:
  - ES-L1
  - GEO
  - HEO
  - SSO

- SO orbit has 0 success rate.



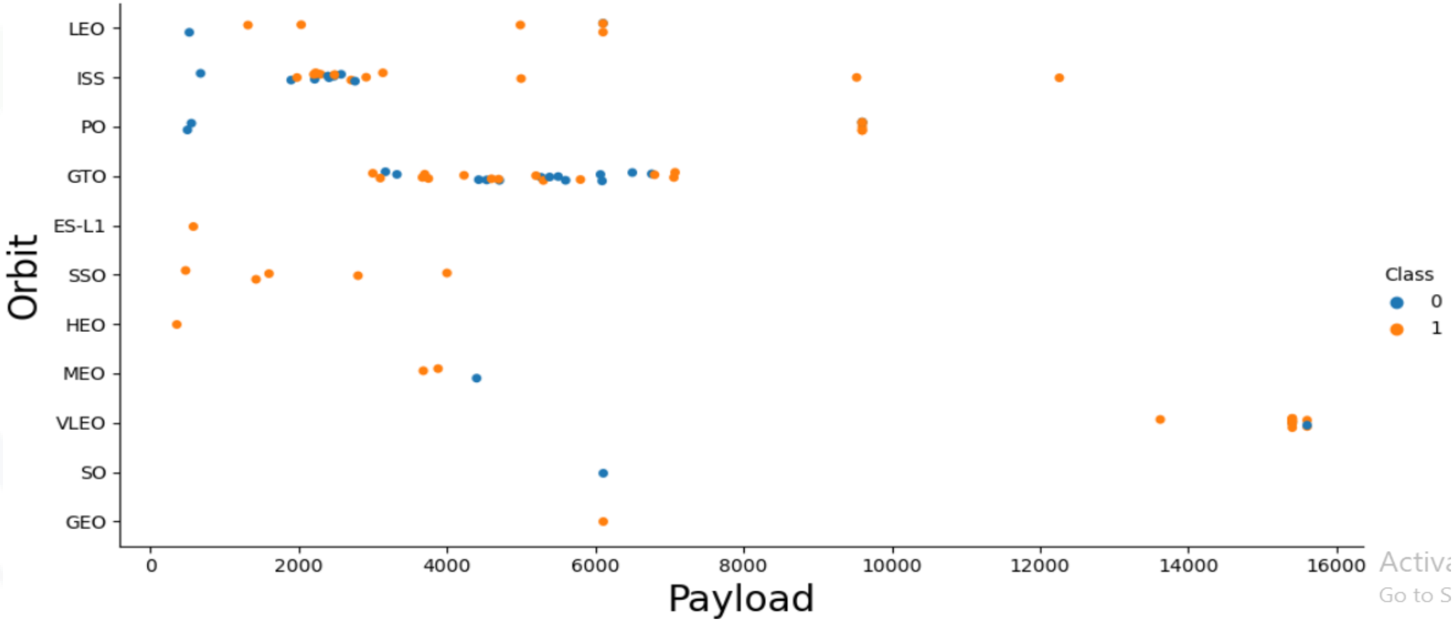
FLIGHT NUMBER vs ORBIT TYPE

- An increased success rate increases success rate is observed for all the orbits.
- SSO and VLEO orbits seem to have come into use recently and seem very promising.



PAYLOAD vs ORBIT TYPE

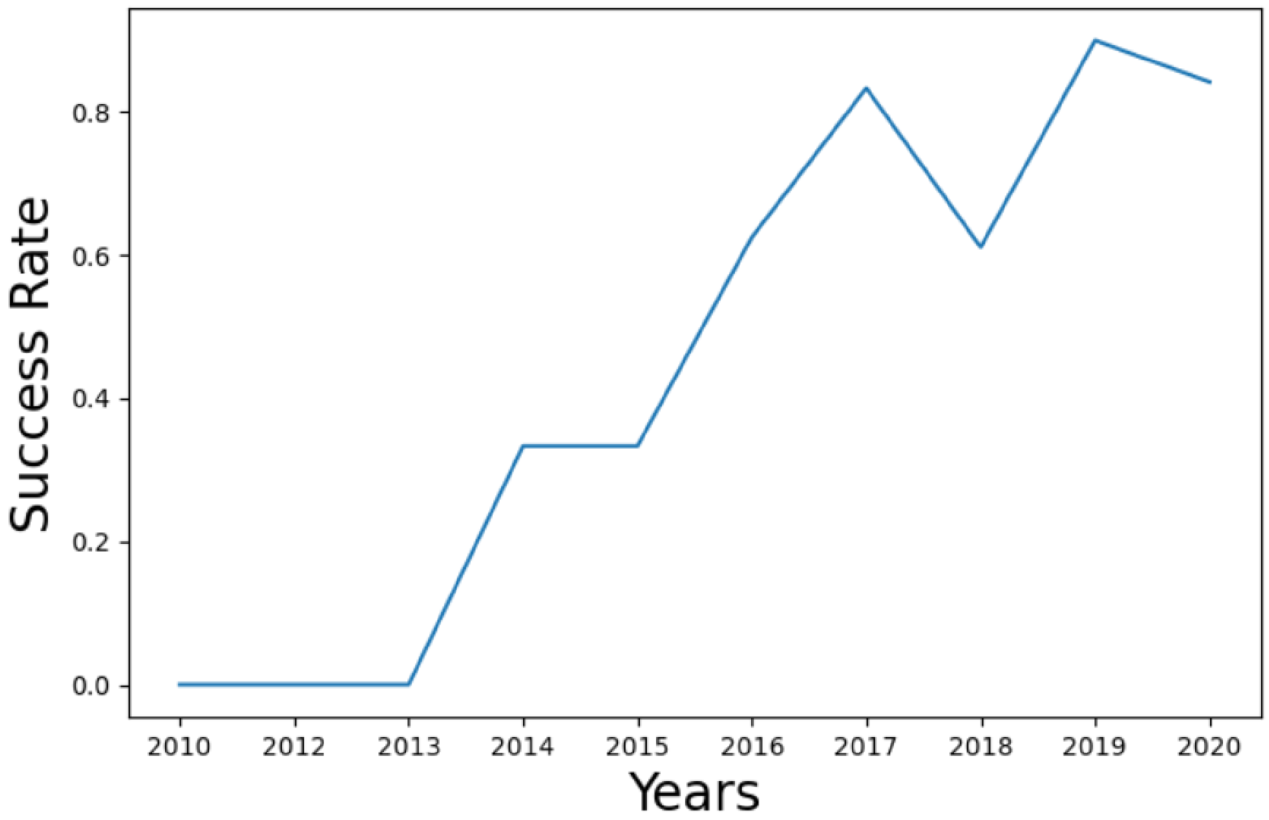
- With heavy payloads the success rate are more for Polar, LEO and ISS orbits.
- For GTO we cannot distinguish this well as the frequency of successful and unsuccessful launches are high.



IBM Developer
SKILLS NETWORK

SUCCESS RATE OVER THE YEARS

- The success rate clearly has improved over the years after the year 2013.



THE LAUNCH SITES

- The unique launch sites are.

```
[7]: sql select distinct Launch_Site from spacextbl

 * sqlite:///my_data1.db
Done.
```

[7]:

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

LAUNCH SITES that begin with 'CCA'

- The launch missions for which the launch site name begins with "CCA".

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

```
[14]: sql select * from spacextbl where Launch_Site like 'CCA%' limit 5
```

```
 * sqlite:///my_data1.db
Done.
```

[14]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing _Outcome
04-06-2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
08-12-2010	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)
22-05-2012	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
08-10-2012	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
01-03-2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

TOTAL PAYLOAD

- Total payload carried by boosters launched by NASA.

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

```
[17]: sql select sum(payload_mass__kg_) from spacextbl where customer = 'NASA (CRS)';

 * sqlite:///my_data1.db
Done.
```

[17]:

sum(payload_mass__kg_)
45596

AVERAGE PAYLOAD

- Average payload carried by boosters version F9 v1.1.

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

[18]:
```
sql select avg(payload_mass__kg_) from spacextbl where Booster_Version = 'F9 v1.1';
```

```
 * sqlite:///my_data1.db
Done.
```

[18]:

avg(payload_mass__kg_)
2928.4

DATE OF THE FIRST SUCCESSFUL LAUNCH

- Date of first successful launch for ground pad.

first_success_gp
2015-12-22

SUCESSFUL BOOSTERS IN DRONE SHIPS

- Names of boosters successful in drone ship with payload mass between 4000 and 6000 kg.

```
[25]: sql select Booster_Versio

 * sqlite:///my_data1.db
Done.
```

[25]:

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

IBM Developer

SKILLS NETWORK

TOTAL SUCCESSFUL AND FAILED LAUNCHES

- Total number of successful and failed launches.

List the total number of successful and failure mission outcomes

```
[26]: sql select Mission_Outcome, count(*) from spacextbl group by Mission_Outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

[26]:

Mission_Outcome	count(*)
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

BOOSTER VERSIONS WITH MAXIMUM PAYLOAD

- Booster versions with highest payload mass.

List the names of the booster_versions which have carriec

```
[27]: sql select distinct(Booster_Version) from spacextbl
```

```
 * sqlite:///my_data1.db
Done.
```

[27]:

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

LAUNCH RECORDS FOR 2015

- Failed landing outcomes in drone ship in year 2015.

[29]:
```
sql select substr(Date, 4, 2) as Month, "Landing _Outcom
```

```
 * sqlite:///my_data1.db
Done.
```

[29]:

Month	Landing _Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

RANK OF LANDING OUTCOMES

- Rank the count of successful landing_outcomes between the date **04-06-2010** and **20-03-2017** in descending order.

landing__outcome	qty
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

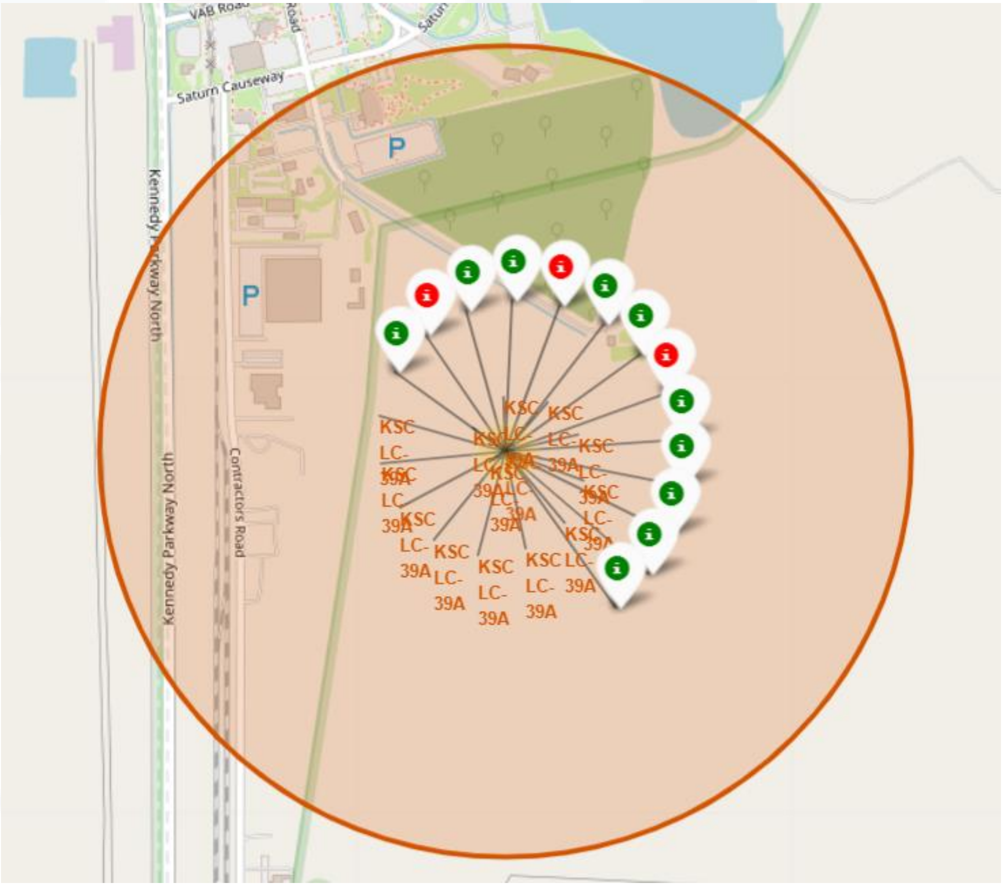
ALL LAUNCH SITES ON THE MAP

- All the launch sites are away from the populated areas (cities) near the coastlines but close enough to sophisticated infrastructure like railways.

LAUNCH OUTCOMES FOR EACH SITE

- This shows the KSC LC-39A launch site.

- The green and red markers denote the successful and failed launch missions respectively.

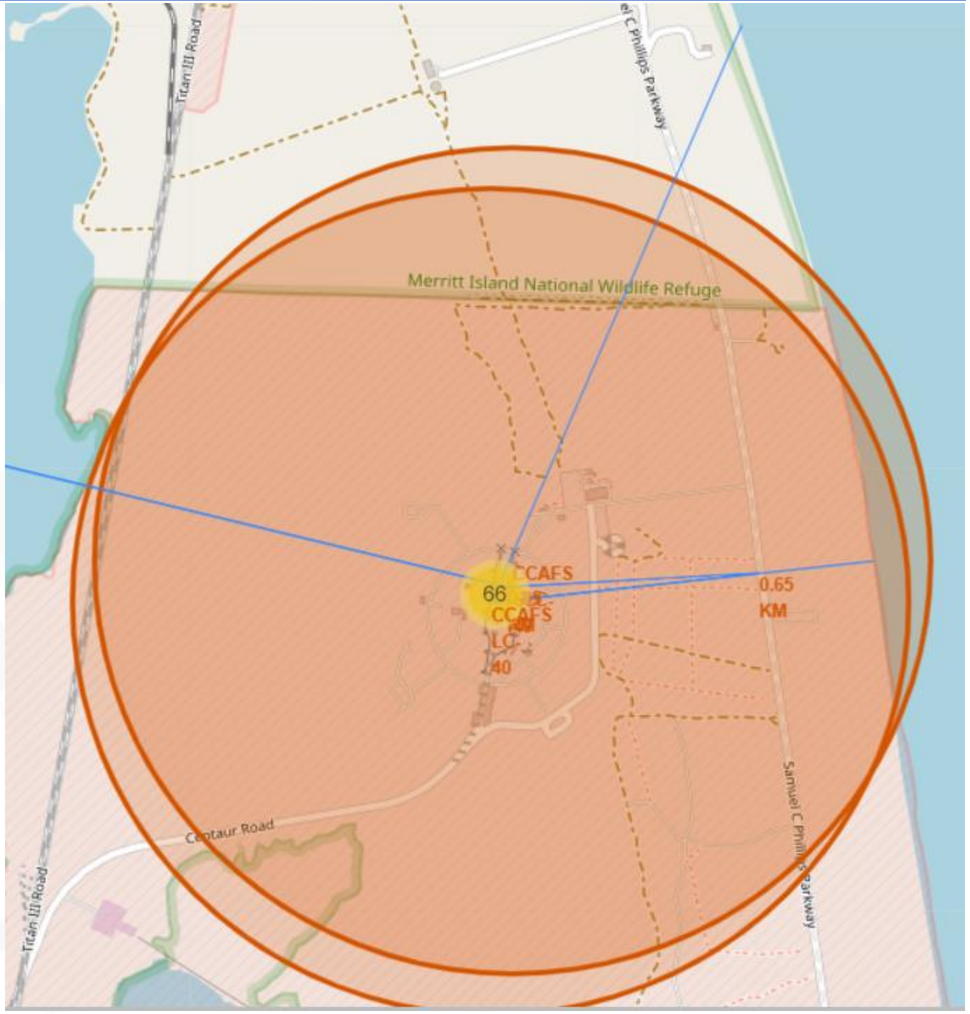


IBM Developer
SKILLS NETWORK

INFRASTRUCTURE AND SAFETY

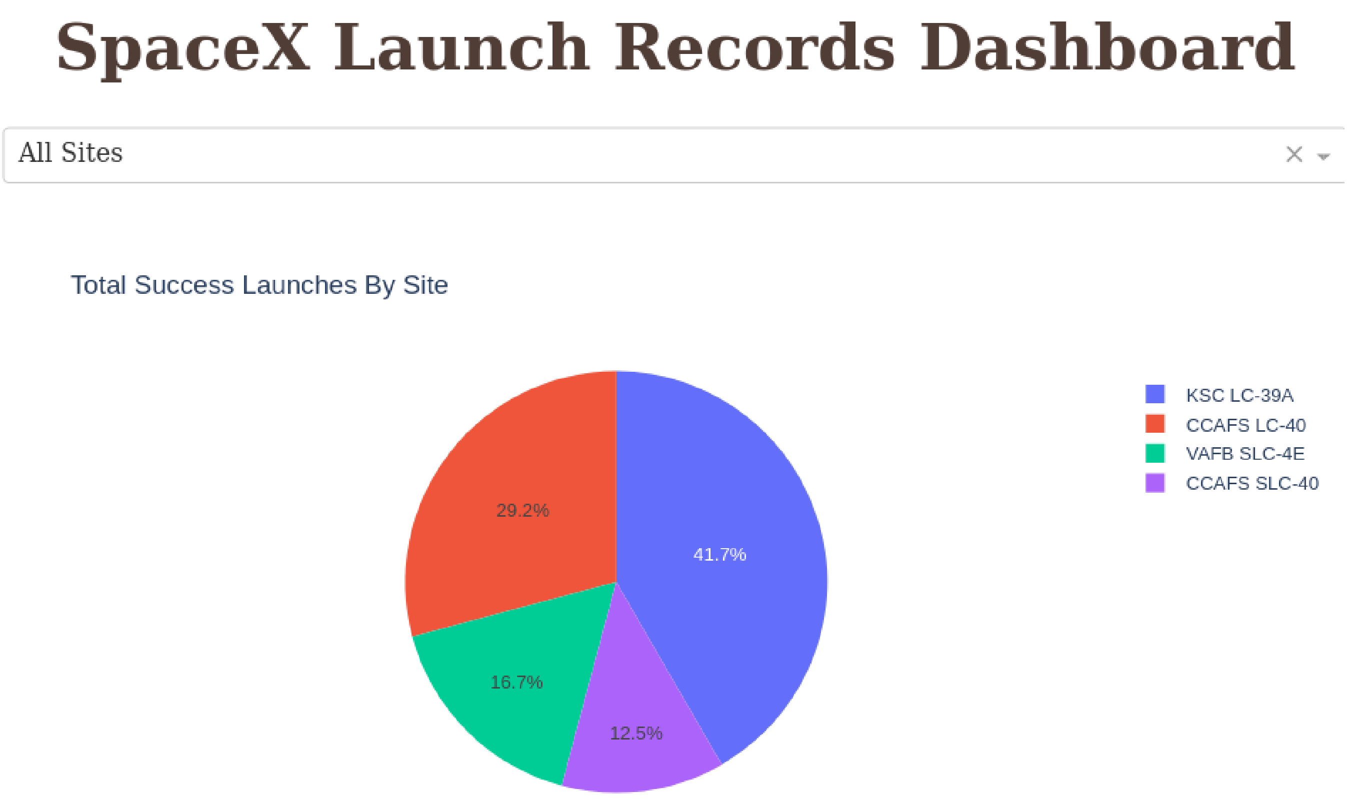
- Launch sites have a sophisticated infrastructure as they have good railways and roads in the vicinity.

- The sites are also far away from the populated cities thus ensuring safety.



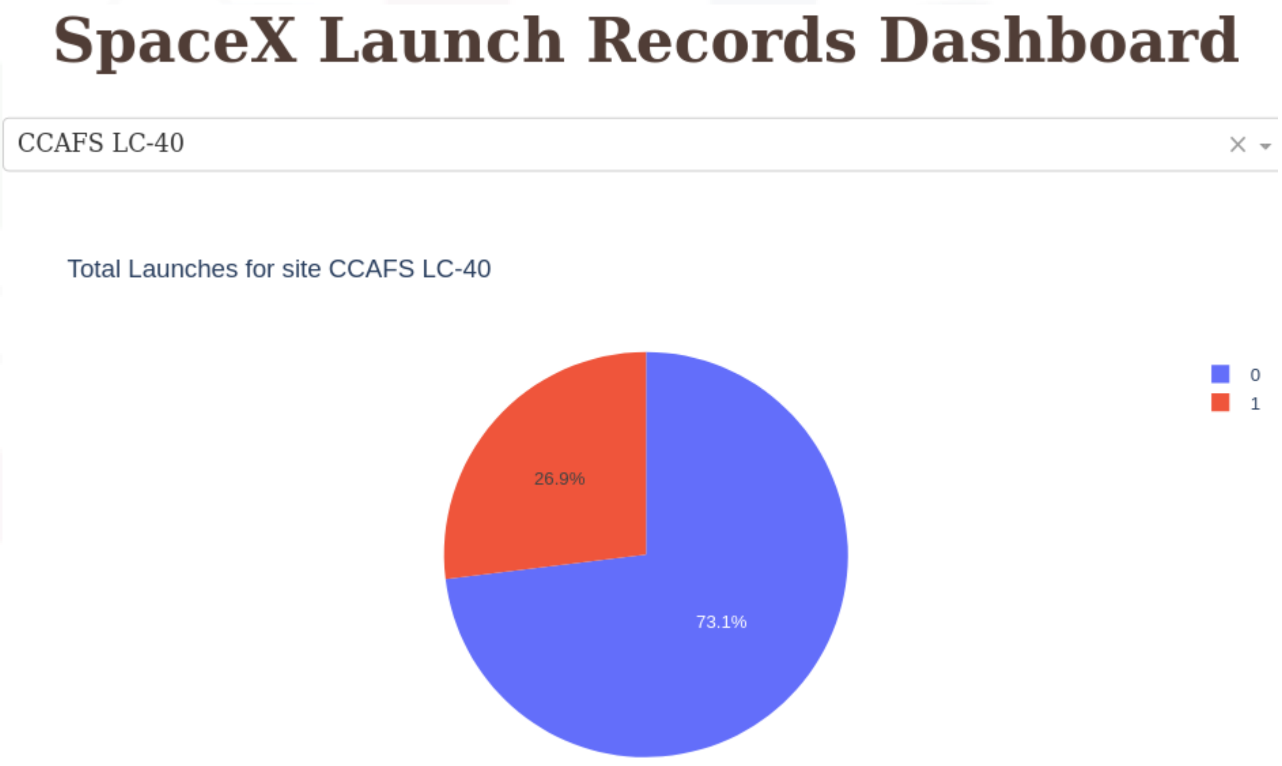
SUCCESSFUL LAUNCHES BY LAUNCH SITES

- The launch sites is an important factor affecting the success of the launch mission.



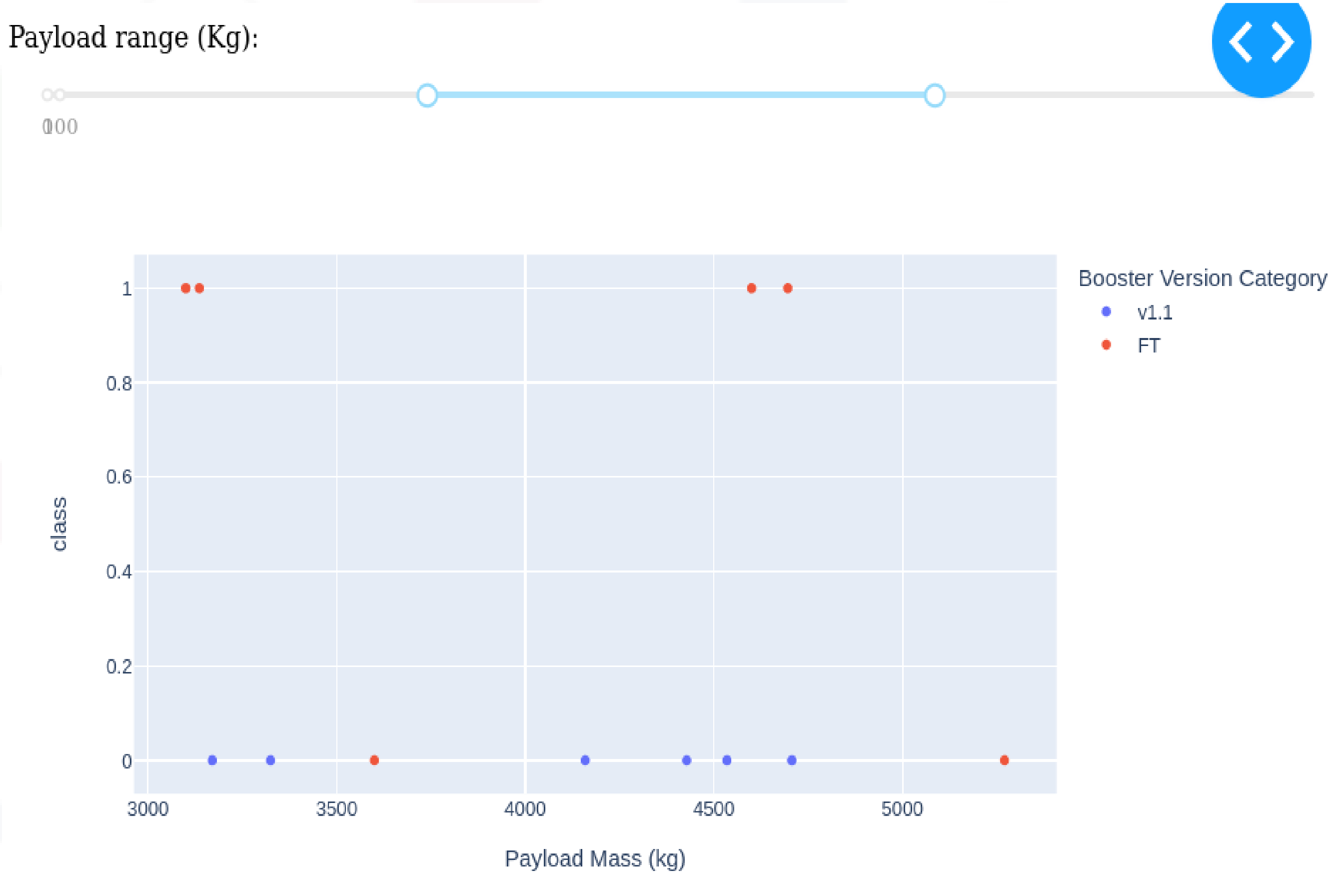
SUCCESS RATE FOR CCAFS LC-40

- CCAFS LC-40 site reports 73.1 % success in all of its mission launches.



PAYLOAD AND LAUNCH OUTCOME RELATION

- Payloads in the range of 3000 to 6000 kg for v1.1 booster result in failures for all launch missions.

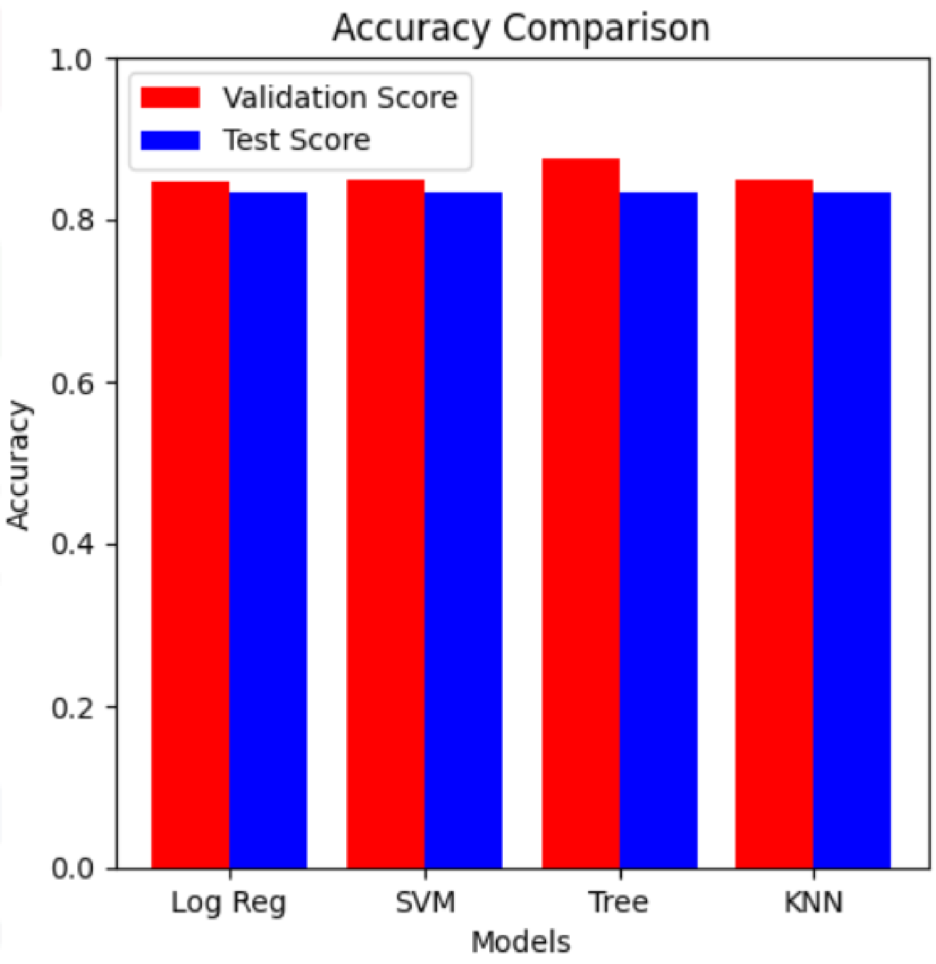


CLASSIFICATION ACCURACY OF EACH MODEL

- Four classification models were used.
    - Logistic Regression
    - Support Vector Machines
    - Decision Trees.
    - K Nearest Neighbors

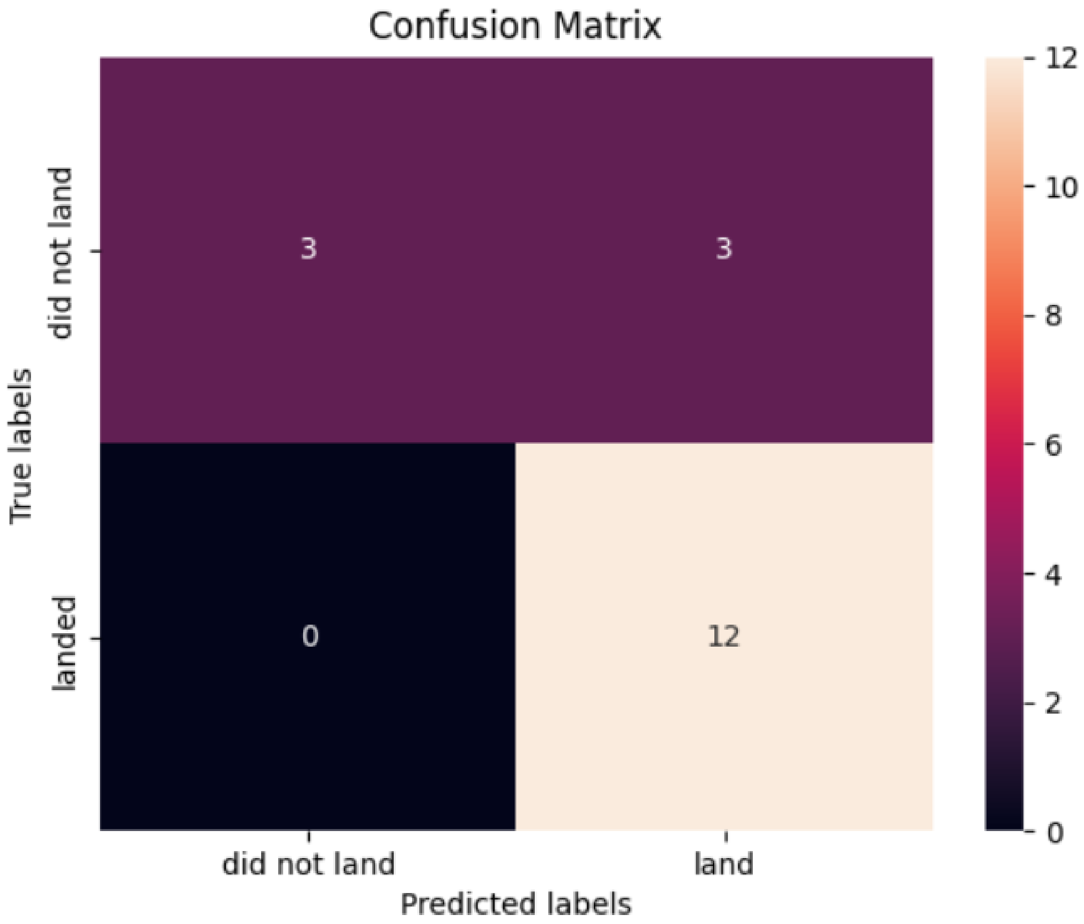
- The test accuracies of each model were the same. 83%.

- The validation accuracy of the Decision Tree model was the highest.



DECISION TREE CLASSIFIER

- The confusion matrix of the Decision Tree shows that the model did a better job in classification.



CONCLUSION

- Multiple data sources were analyzed throughout the process to make a final conclusion.
- The site KSC LC-39A is the best site with the highest success rate for launch missions.
- Launches above 6000 kg of payload mass are less risky.
- The Decision Tree classifier was the best model in predicting the success and failure of the launches and thus increasing the profits.

APPENDIX

- All the notebooks have been uploaded on GitHub – IBM Data Science Final Capstone Project.