ANALYZING LAUNCH REUSABILITY FOR SPACEX

HUA CHEN

12/09/2023

OUTLINE

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

EXECUTIVE SUMMARY

Methodological Overview

- **Data Procurement:** API extraction, SQL querying, and web scraping
- **Data Preparation & Analysis:** for success/failure outcomes & Analyzed key metrics
- **Geographical Insights:** scrutinize launch site success vs. geographical landmarks, noting historical wins and payload ranges
- **Predictive Modeling:** Construct and compare models

EXECUTIVE SUMMARY

Key Findings

- Insights from data analysis enhanced with interactive visuals for clarity
- Post evaluation, select the best model for predicting landing outcomes

INTRODUCTION

Background

In the race to democratize space travel, SpaceX stands out with its cost-effective rocket launches, thanks in large part to the groundbreaking reuse of the Falcon 9's first stage. While traditional launches from other providers can surge beyond $165 million, SpaceX has streamlined this to a mere $62 million. The key to these savings? Successfully landing and reusing the rocket's first stage. Our objective is to harness public data and machine learning to predict the reusability of this crucial stage, potentially offering insights for potential competitors.

INTRODUCTION

Problems

1. Assessing the impact of variables such as payload mass, launch site, number of flights, and orbits on first-stage landing outcomes.

2. Monitoring the success rate of landings over time.

3. Determining the optimal predictive model for landing success.

4. Delving into the primary factors and conditions that influence a rocket's landing success and distinguishing between successful and failed attempts.

5. Understanding the comprehensive effects of rocket variables on SpaceX's landing outcomes and optimizing conditions for success.

METHODOLOGY

Data Collection

- Utilize SpaceX REST API
- Employ web scraping techniques

Data Wrangling

- Implement one-hot encoding for ML-ready data

Exploratory Data Analysis(EDA)

- EDA with SQL
- EDA with Visualization

Interactive Visual Analytics

- Mapping Insights with Folium
- Dynamic Visualizations with Plotly Dash

Predictive Analysis

- Build classification models
- Evaluate model performance

Utilize SpaceX REST API

METHODOLOGY

Data Collection

1. **Request data from SpaceX API**

```
spacex_url="https://api.spacexdata.com/v4/launches/past"
```

```
response = requests.get(spacex_url)
```

2. **Transform response into .json format and convert the result into dataframe**

```
# Use json_normalize meethod to convert the json result into a dataframe
data_json = response.json()
data = pd.json_normalize(data_json)
```

Employ web scraping techniques

METHODOLOGY

Data Collection

1. Request data from Wiki page

```
static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"
# use requests.get() method with the provided static_url
# assign the response to a object
response = requests.get(static_url)
response.status_code
```

2. Using BeautifulSoup to extract data

```
soup = BeautifulSoup(response.content, 'html.parser')
# Use the find_all function in the BeautifulSoup object, with element type `table`
# Assign the result to a list called `html_tables`
html_tables = soup.find_all('table')
# Let's print the third table and check its content
first_launch_table = html_tables[2]
print(first_launch_table)
```

Employ web scraping techniques

3. Generate dataframe from launch HTML tables.

```
launch_dict= dict.fromkeys(column_names)

# Remove an irrelvant column
del launch_dict['Date and time ( )']

# Let's initial the launch_dict with each value to be an empty list
launch_dict['Flight No.'] = []
launch_dict['Launch site'] = []
launch_dict['Payload'] = []
launch_dict['Payload mass'] = []
launch_dict['Orbit'] = []
launch_dict['Customer'] = []
launch_dict['Launch outcome'] = []
# Added some new columns
launch_dict['Version Booster']=[]
launch_dict['Booster landing']=[]
launch_dict['Date']=[]
launch_dict['Time']=[]
```

```
df=pd.DataFrame(launch_dict)
df.head()
```

	1	2	3	4	5	6	7	8	9	10	...	112	113	11
Flight No.	1	2	3	4	5	6	7	8	9	10	...	112	113	11
Date	4 June 2010	8 December 2010	22 May 2012	8 October 2012	1 March 2013	29 September 2013	3 December 2013	6 January 2014	18 April 2014	14 July 2014	...	24 March 2021	7 April 2021	23 Apr 202
Time	18:45	15:43	07:44	00:35	15:10	16:00	22:41	22:06	19:25	15:15	...	08:28	16:34	9:4
Version Booster	F9 v1.0B0003.1	F9 v1.0B0004.1	F9 v1.0B0005.1	F9 v1.0B0006.1	F9 v1.0B0007.1	F9 v1.1B1003	F9 v1.1	F9 v1.1	F9 v1.1	F9 v1.1	...	F9 B5B1060.6	F9 B5 △	F B5B1061.
Launch Site	CCAFS	CCAFS	CCAFS	CCAFS	CCAFS	VAFB	CCAFS	CCAFS	Cape Canaveral	Cape Canaveral	...	CCSFS	CCSFS	KS

rows x 121 columns

METHODOLOGY

Data Collection

METHODOLOGY

Data Wrangling

Implement one-hot encoding for ML-ready data

1. Prepare Set for binary label for Landing

```
# landing_outcomes = values on Outcome column
landing_outcomes = df['Outcome'].value_counts()
landing_outcomes
for i,outcome in enumerate(landing_outcomes.keys()):
    print(i,outcome)
bad_outcomes=set(landing_outcomes.keys()[[1,3,5,6,7]])
bad_outcomes
```

```
{'False ASDS', 'False Ocean', 'False RTLS', 'None ASDS', 'None None'}
```

2. Create binary label for landing outcome

```
# landing_class = 0 if bad_outcome
# landing_class = 1 otherwise
landing_class = [0 if outcome in bad_outcomes else 1 for outcome in df['Outcome']]
df['Class']=landing_class
df[['Class']].head(8)
```

METHODOLOGY

Exploratory Data Analysis(EDA)

EDA with SQL

Queries to display from Dataset

- Unique launch site names
- First 5 records with launch sites starting with 'CCA'
- Total payload mass for boosters launched by NASA (CRS)
- Average payload mass by booster version F9v1.1

Queries to list from Dataset

- Date of the first successful ground pad landing
- Booster names with successful drone ship landings, payload between 4,000 and 6,000
- Total count of successful and failed missions
- Booster versions carrying the maximum payload
- Failed landings on drone ships in 2015 with booster version and launch site details
- Landing outcomes count between 2010-06-04 and 2017-03-20 (in descending order)

METHODOLOGY

Exploratory Data Analysis(EDA)

EDA with Visualization

Scatter Plot

- Payload vs Flight Number
- Flight Number vs Launch Site
- Payload vs Launch Site
- Flight Number vs Orbit Type
- Payload vs Orbit Type

Scatter plots illustrate the interrelation of features. After identifying trends in the graph, we can easily forecast the elements that increase the likelihood of success in both results and landings. If a connection is observed among the variables, they can be valuable for machine learning applications.

METHODOLOGY

Exploratory Data Analysis(EDA)

EDA with Visualization

Bar Plot

- Success Rate vs Orbit Type

Bar plot effectively represent comparisons the Success Rate between different categories of Orbit Type, and it can clearly identify which orbits possess the highest success rates

Line Plot

- Success Rate yearly trend

Line plot clearly depict the annual trend in success rate

Mapping Insights with Folium

METHODOLOGY

Interactive Visual Analytics

1. Interactive Leaflet Map

- Utilize Folium for an enhanced, interactive map visualization
- Highlight launch sites with circle markers based on coordinates

2. Launch Site Markers

- Blue circle at NASA Johnson Space Center with a popup displaying its name
- Red circles for all other launch sites with popups showcasing their names

METHODOLOGY

Interactive Visual Analytics

Mapping Insights with Folium

3. Outcome Visualization

- Green markers for successful launches; Red for unsuccessful ones
- Allows a quick view of launch sites with the highest success rates

4. Launch Site Proximity Analysis

- Display distances from launch site CCAFS SLC-40 to key landmarks like coastlines, railways, highways, and cities.

METHODOLOGY

Interactive Visual Analytics

Dynamic Visualizations with Plotly Dash

1. Launch Site Selection

- Dropdown option for users to view all or specific launch sites

2. Launch Success Overview

- Pie chart representation of both successful and failed launches, showcasing the success percentage

3. Payload Insights

- A slider feature to adjust and view data within specific payload mass ranges

- Scatter chart detailing the interplay between payload mass and launch success across different booster versions

METHODOLOGY

Interactive Visual Analytics

Dynamic Visualizations with Plotly Dash

4. Deep Dive into Launch Sites

- Pie chart dissecting the success rate specific to each launch site

- Scatter plot highlighting the correlation between payload mass and success, further categorized by booster versions

METHODOLOGY

Predictive Analysis

Machine Learning Workflow

Model Construction

↓

Model Evaluation

↓

Optimal Model Determination

METHODOLOGY

Predictive Analysis

Model Construction

- Initiate with feature-engineered data and convert to numpy arrays
- Standardize data for uniformity
- Segregate data into training and testing subsets
- Overview of chosen machine learning algorithms
- Utilize GridSearchCV for parameter tuning and model training

```
from sklearn.model_selection import train_test_split

transform = preprocessing.StandardScaler()
X = transform.fit_transform(X)

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```
logreg_cv = GridSearchCV(lr, parameters1, cv=10)
logreg_cv.fit(X_train, Y_train)
```

METHODOLOGY

Predictive Analysis

Model Evaluation

- Assess accuracy metrics for all models
- Extract optimal hyperparameters from each algorithm
- Visual representation using confusion matrices

```
yhat=logreg_cv.predict(X_test)
plot_confusion_matrix(Y_test,yhat)
```

METHODOLOGY

Predictive Analysis

Optimal Model Determination

- Best model recognized based on out-of-sample accuracy

```
test_accuracy = svm_cv.score(X_test, Y_test)
print("Test Accuracy: ", test_accuracy)
```

METHODOLOGY

Predictive Analysis

Optimal Model Determination

- Best model recognized based on Accuracy, Efficiency, Precision and recall

METHODOLOGY

Predictive Analysis

Techniques

- Employ StandardScaler for data standardization
- Data split using train_test_split for model validation
- Parameter optimization through GridSearchCV across multiple algorithms: Logistic Regression, SVM, Decision Tree, K-Nearest Neighbor
- Utilize confusion matrix to visualize true positives, false negatives, true negatives, and false positives
- Use Accuracy metrics like accuracy_score, precision_score, recall _score and compare the training accuracy, test accuracy and model efficiency to finalize the best model

RESULTS

EDA Results

- KSC LC-39A stands out with topmost success among landing locations
- Perfect success observed in ES-L1, GEO, HEO, and SSO orbits
- Trend of increasing launch success as time progresses

Interactive Analytics Results

- Most launch sites are near the equator and coast

Predictive Analytics Results

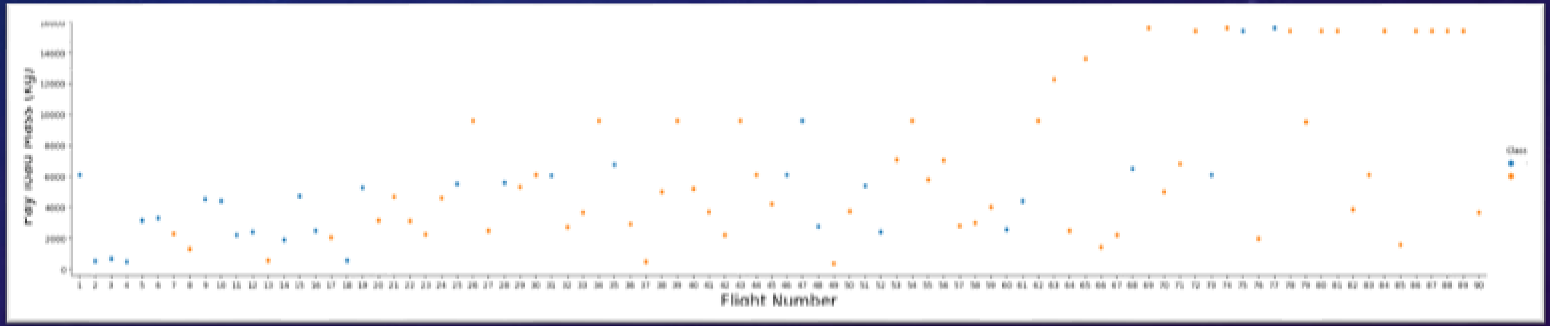
- All models produced same results in the confusion matrix
- The Logistic Regression outperformed the other models

RESULTS

EDA Results

Flight Number vs. Payload Mass

- As the flight number rises, the likelihood of a successful first-stage landing also increases
- However, a heavier payload appears to reduce the chances of the first stage's successful return

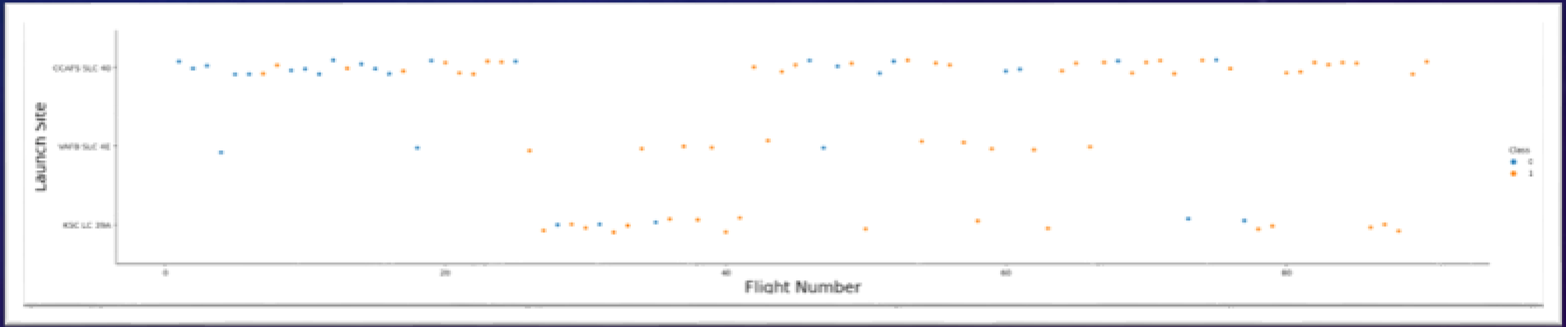


RESULTS

EDA Results

Flight Number vs. Launch Site

- Newer launches are more successful
- The CCAFS SLC 40 site accounts for about 50% of all launches
- Both VAFB SLC 4E and KSC LC 39A are characterized by their higher success frequencies

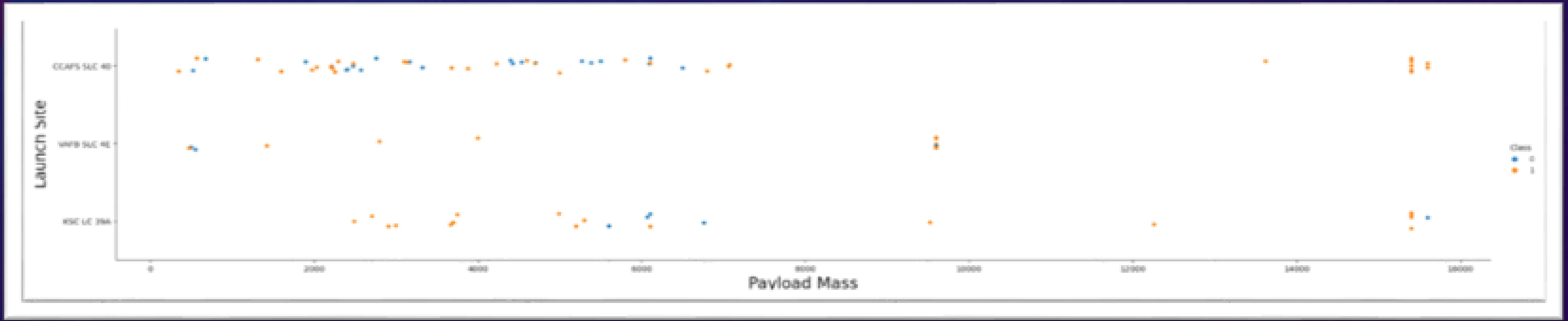


RESULTS

EDA Results

Payload Mass vs. Launch Site

- No rockets with a payload mass exceeding 10,000 kg were launched from VAFB-SLC
- Success rates generally increase with heavier payload masses
- Launches carrying more than 7,000 kg often achieve success
- Every launch from KSC LC 39A weighing below 5,500 kg has been successful
- VAFB SKC 4E hasn't attempted any launch surpassing roughly 10,000 kg

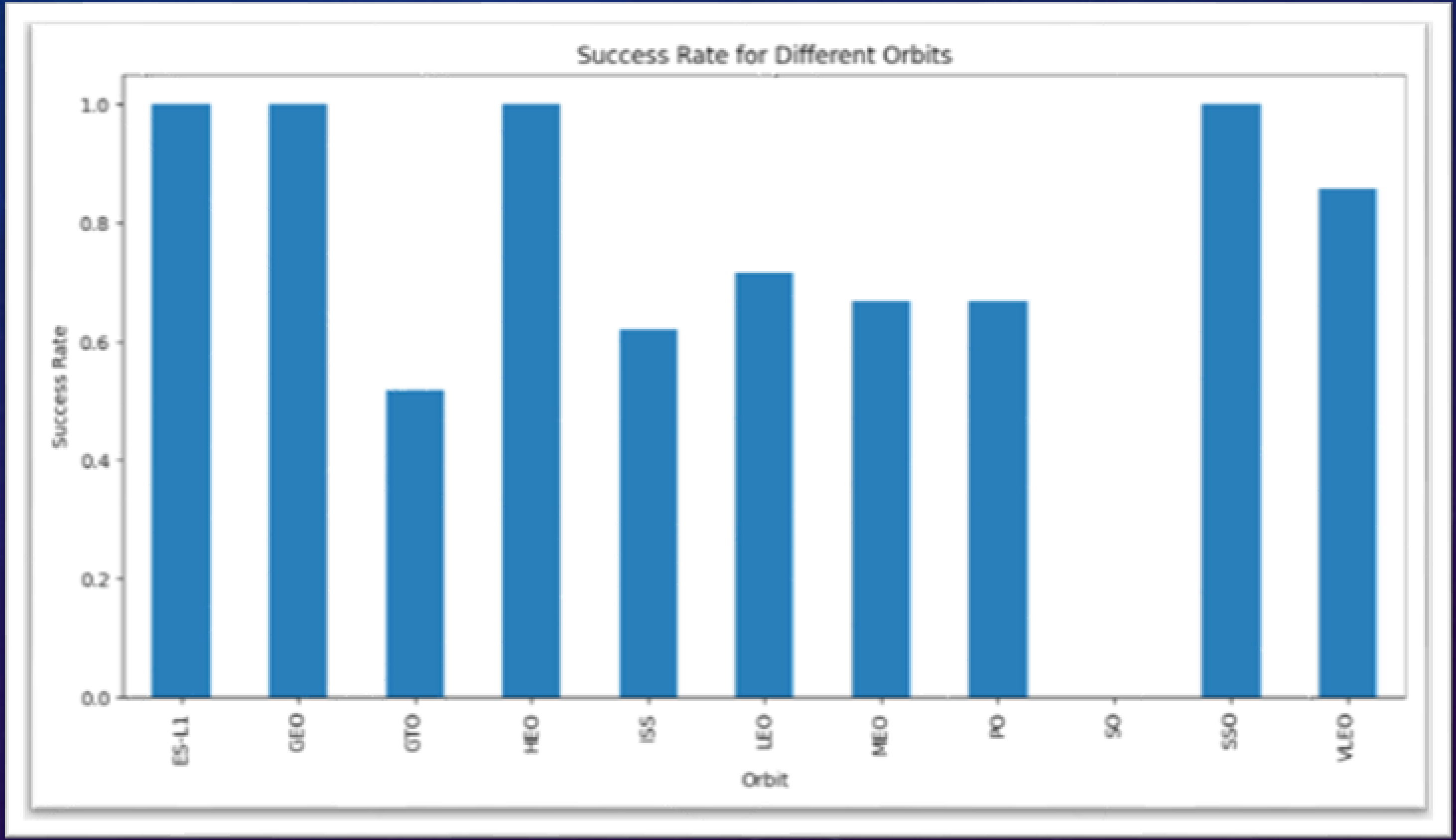


RESULTS

EDA Results

Orbit Type vs. Success Rate

- Full Success (100%): ES-L1, GEO, HEO, SSO
- Moderate Success (50%-80%): GTO, ISS, LEO, MEO, PO
- No Success (0%): SO

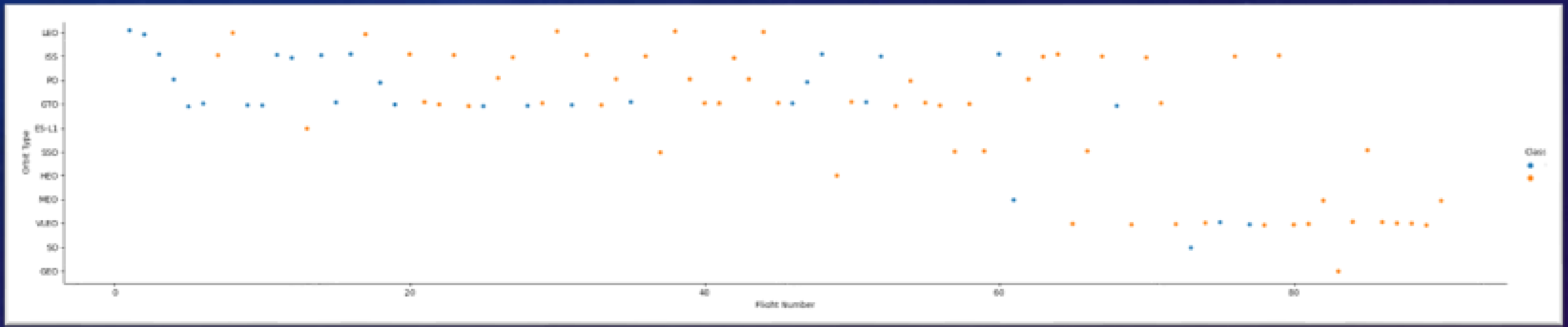


RESULTS

EDA Results

Flight Number vs. Orbit Type

- A rising trend in success rate is observed as the number of flights increases, especially evident in the LEO orbit
- Conversely, the GTO orbit doesn't adhere to this increasing pattern

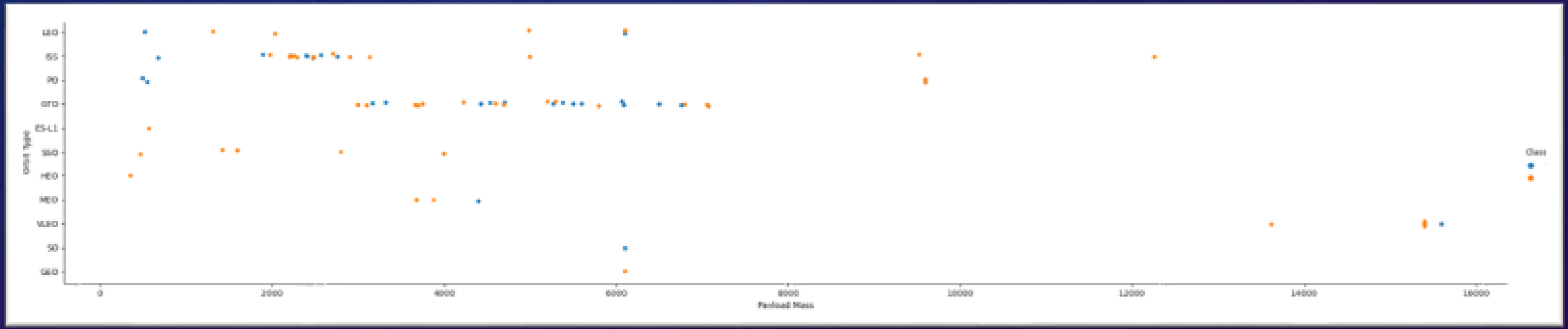


RESULTS

EDA Results

Payload Mass vs. Orbit Type

- Heavier payloads tend to have a higher success rate for Polar, LEO, and ISS orbits
- For GTO orbit, there's a mix of both successful and unsuccessful landings, making it harder to draw a clear correlation

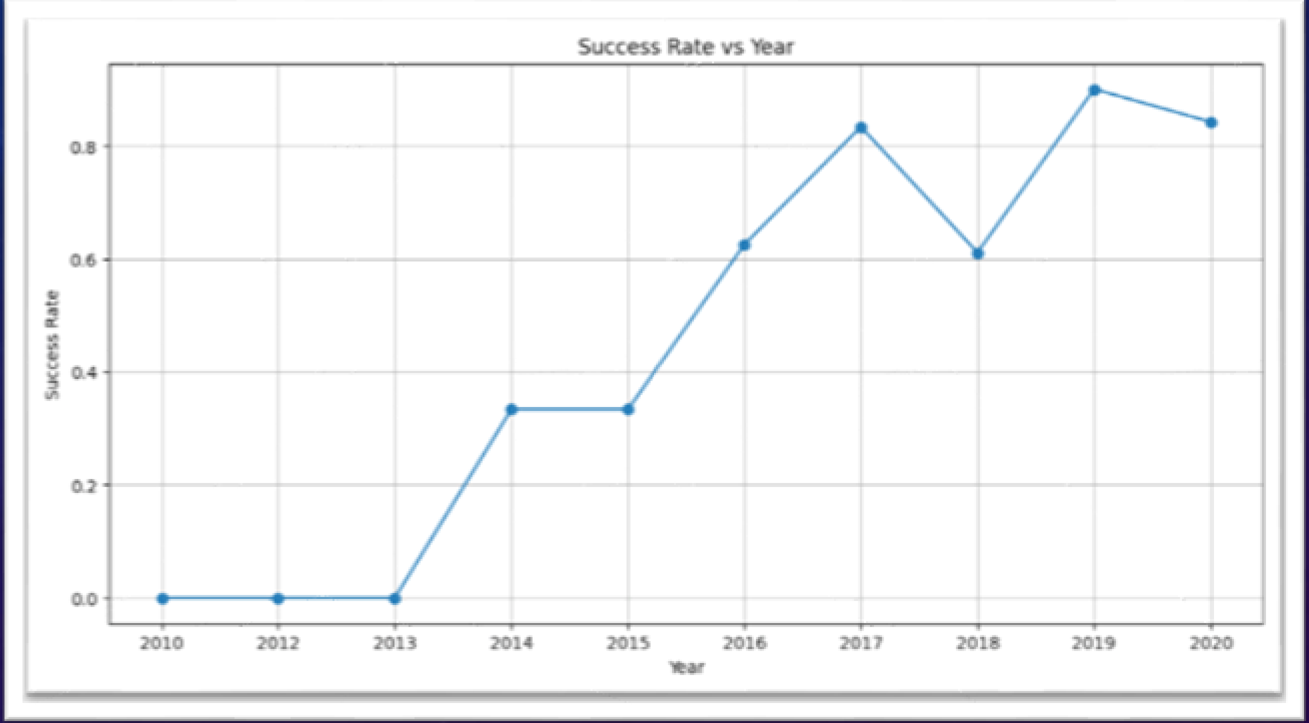


RESULTS

EDA Results

Success Rate vs. Year

- The success rate has generally risen, although there was a minor decline after 2019.



RESULTS

EDA Results

Launch Site Information

- Names of the unique Launch Sites

```
%sql SELECT DISTINCT "Launch_Site" FROM SPACEXTABLE
 * sqlite:///my_data1.db
Done.
```

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

RESULTS

EDA Results

Launch Site Information

- 5 Records with Launch Site begin with 'CCA'

```
%sql SELECT * FROM SPACEXTABLE WHERE "Launch_Site" LIKE 'CCA%' LIMIT 5
* sqlite:///my_data1.db
Done.
```

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

RESULTS

EDA Results

Payload Mass Information

- Total 45,596 kg carried by boosters launched by NASA (CRS)

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Customer" = 'NASA (CRS)'
 * sqlite:///my_data1.db
Done.
SUM("PAYLOAD_MASS__KG_")
45596
```

- Average 2,928.4 kg carried by booster version F9 v1.1

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXTABLE WHERE "Booster_Version" = 'F9 v1.1
 * sqlite:///my_data1.db
Done.
AVG("PAYLOAD_MASS__KG_")
2928.4
```

RESULTS

EDA Results

Landing Outcome Information

- First Successful Landing in Ground Pad

```
%sql SELECT MIN("Date") FROM SPACEXTABLE WHERE "Landing_Outcome" = 'Success (ground pad)

 * sqlite:///my_data1.db
Done.

MIN("Date")
2015-12-22
```

RESULTS

EDA Results

Landing Outcome Information

- Failed Landing in Drone Ship in 2015

```
%sql SELECT substr(Date, 6, 2) as month, "Landing_Outcome", "Booster_Version", "Launch_Site" \
FROM SPACEXTABLE \
WHERE "Landing_Outcome" = 'Failure (drone ship)' AND substr(Date,1,4)='2015';
```

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

month	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

RESULTS

EDA Results

Landing Outcome Information

- Total number of successful and failure mission outcomes

```
%sql SELECT "Mission_Outcome", COUNT(*) FROM SPACEXTABLE GROUP BY "Mission_Outcome";
```

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

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

RESULTS

EDA Results

Landing Outcome Information

- Landing outcomes count between 2010-06-04 and 2017-03-20 (in descending order)

```
%sql SELECT "Landing_Outcome", COUNT(*) as count \
FROM SPACEXTABLE \
WHERE Date BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY "Landing_Outcome" \
ORDER BY count DESC;
```

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

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

RESULTS

EDA Results

Boosters Information

- Boosters successful on drone ships with a payload mass between 4000 and 6000

```
sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "Landing_Outcome" LIKE 'Success (drone ship)' AND "PAYLOAD_MASS__KG_" > 4000 AND "PAYLOAD_MASS__KG_" < 6000
 * sqlite:///my_data1.db
Done.
```

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

RESULTS

EDA Results

Boosters Information

- Booster versions with the highest payload mass

```
%sql SELECT "Booster_Version" FROM SPACEXTABLE WHERE "PAYLOAD_MASS__KG_" = (SELECT MAX("PAYLOAD_MASS__KG_") FROM SPACEXTABLE);
 * sqlite:///my_data1.db
Done.
```

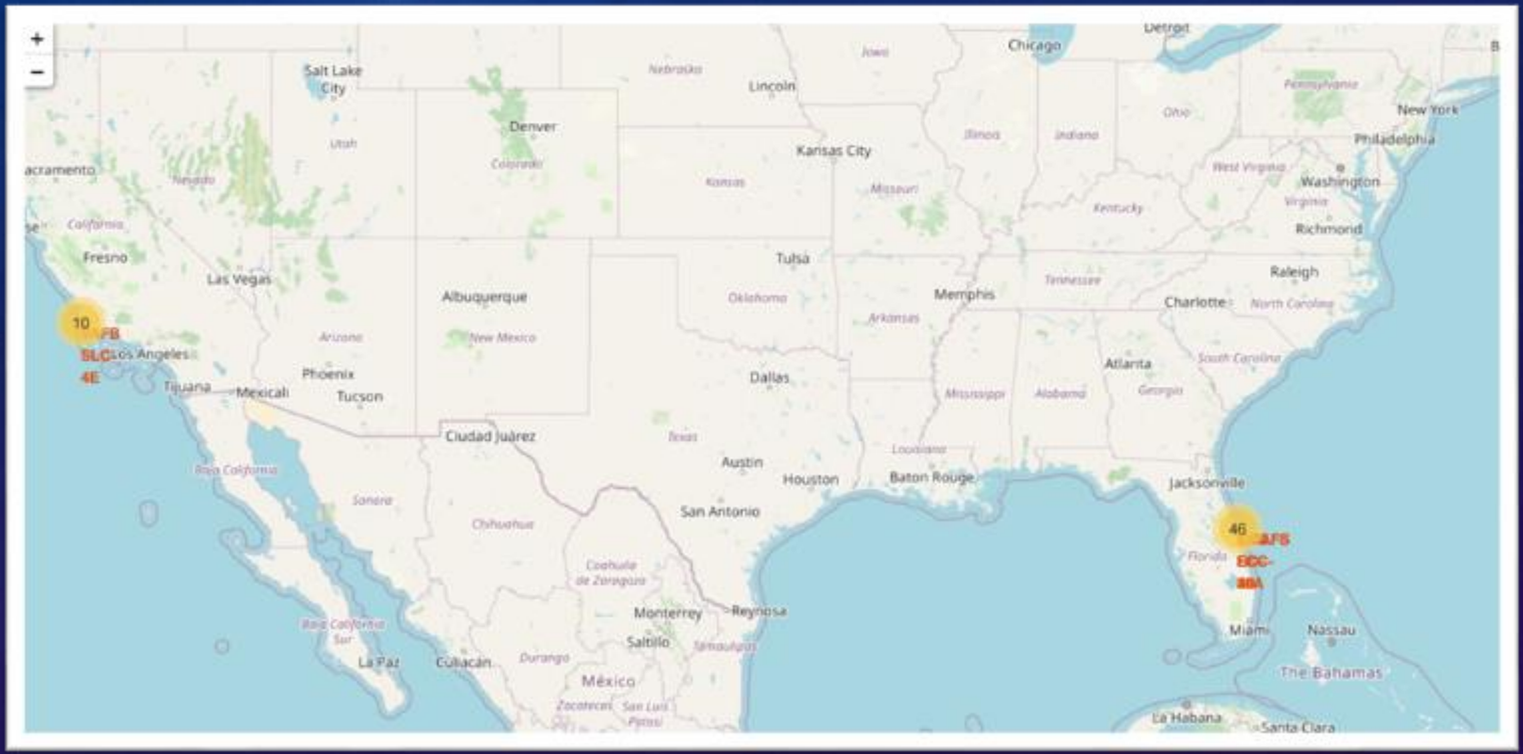
Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

RESULTS

Interactive Analytics Results

Location of Launch Sites

- All launch sites are located in Florida or California, and are close to the equator line and coast



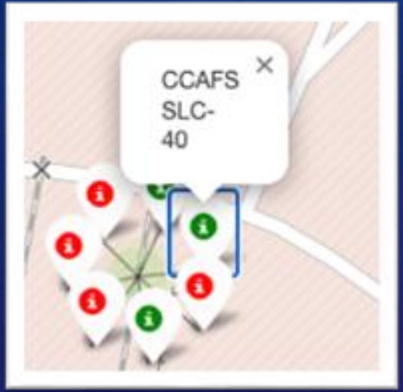
RESULTS

Interactive Analytics Results

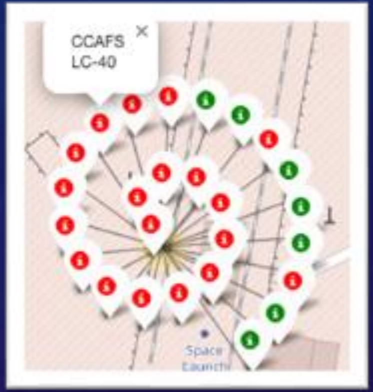
Outcomes of Launch Sites

- At each launch site, Green markers for successful launches Red markers for unsucessful launches:

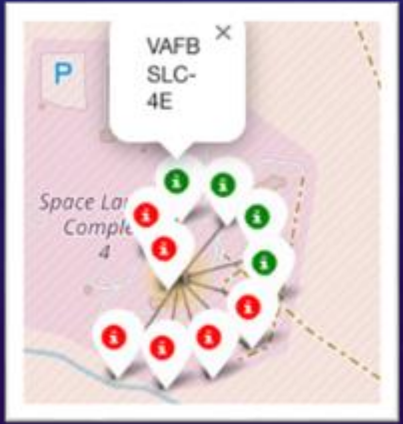
CCAFS SLC-40



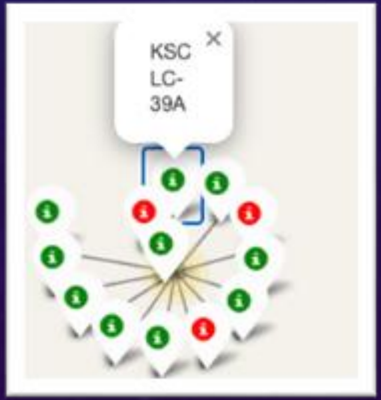
CCAFS LC-40



VAFB SLC-4E



KSC LC-39A



RESULTS

Interactive Analytics Results

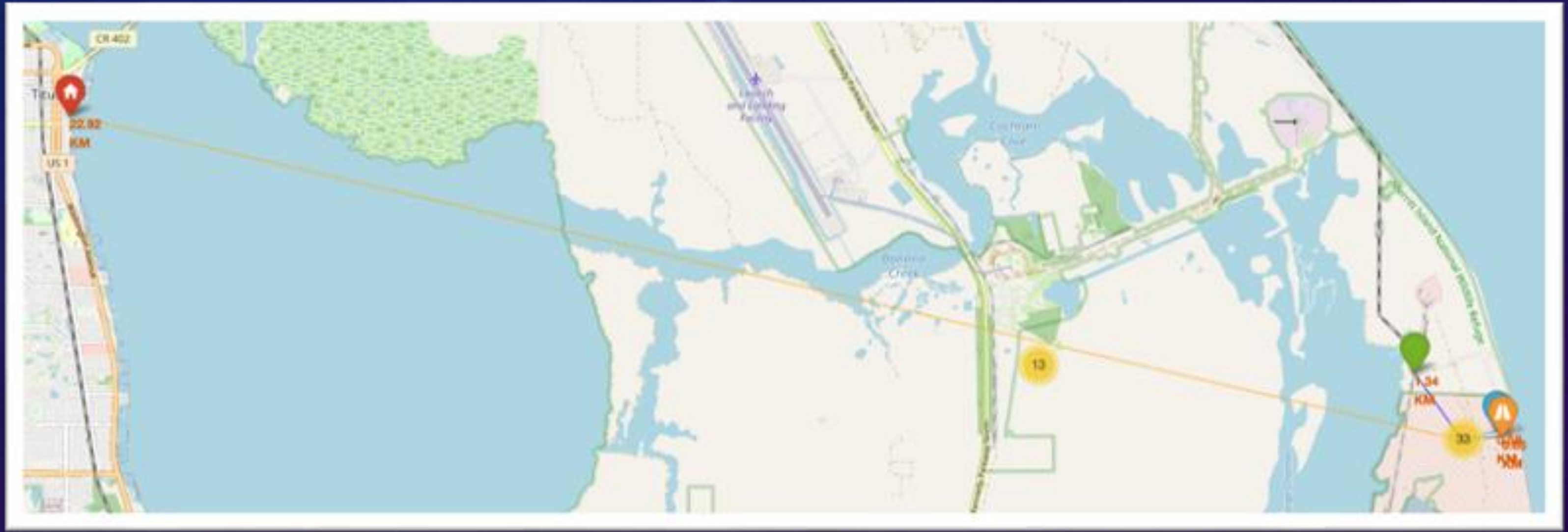
Proximities of Launch Site CCAFS LC-40

Close to the railway, highway and coastline

- 0.58 KM to the closest coastline
- 0.65 KM to the closest highway
- 1.34 KM to the closest railway

Keep away from nearest city

- 22.92 KM to City Titusville



RESULTS

Interactive Analytics Results

Proximities of Launch Site CCAFS SLC-40

Close to the railway, highway and coastline

- 0.51 KM to the closest coastline
- 0.60 KM to the closest highway
- 1.29 KM to the closest railway

Keep away from nearest city

- 22.95 KM to City Titusville

RESULTS

Interactive Analytics Results

Proximities of Launch Site VAFB SLC-4E

Close to the railway, highway and coastline

- 1.33 KM to the closest coastline
- 1.26 KM to the closest railway

Keep away from highway and nearest city

- 14.00 KM to City Lompoc
- 14.91 KM to the closest highway

RESULTS

Interactive Analytics Results

Success Count for all launch sites

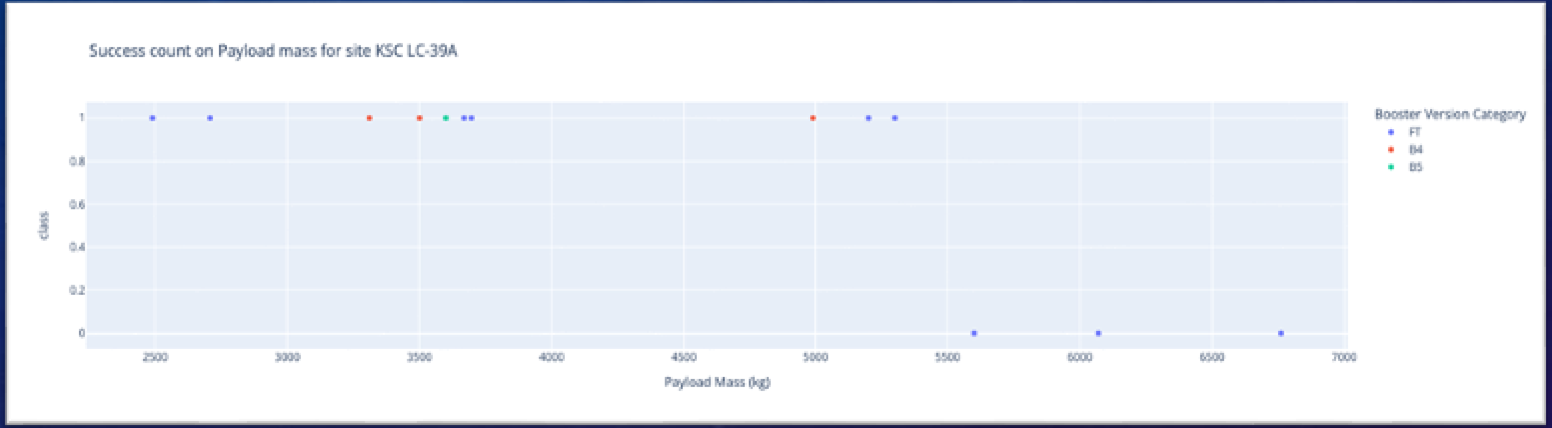
KSC LC-39A has the most successful launches among launch sites with the success rate 76.9%

RESULTS

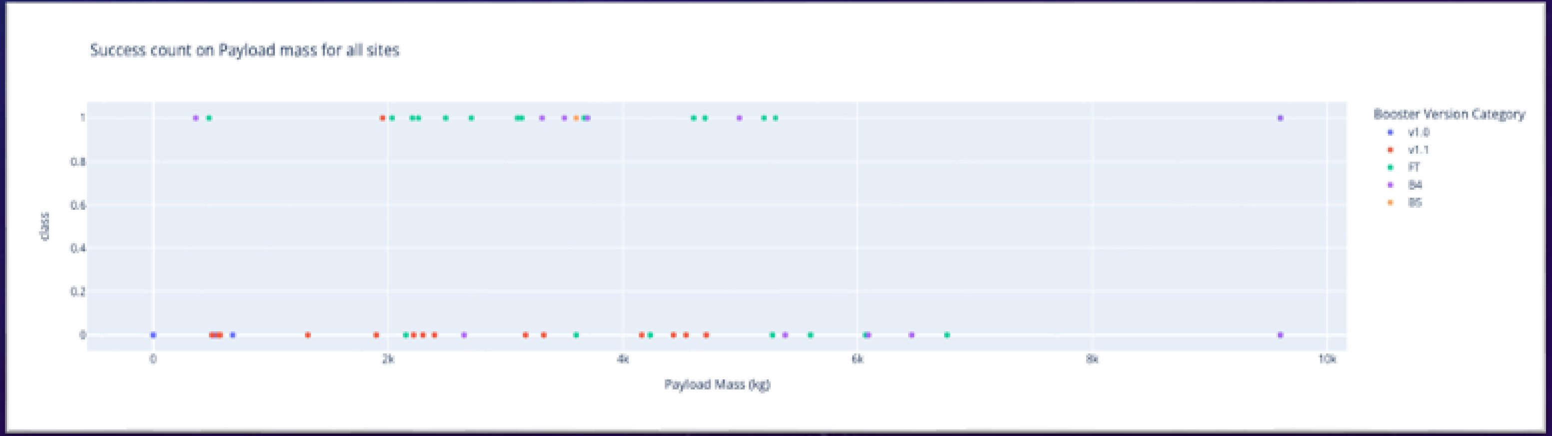
Interactive Analytics Results

Success Count for Payload mass

For site KSC LC-39A the Boosters with the payloads between 2400 KG and 5500 KG have the highest success rate



For all sites the Boosters with the payloads between 1900 KG and 5300 KG have the highest success rate



RESULTS

Predictive Analytics Results

Accuracy of Models

- All the models(Logistic Regression, SVM, KNN, Decision Tree) Have the same Test Accuracy
- Logistic Regression, SVM, KNN Have very similar Training Accuracy
- Decision Tree has higher Training Accuracy in compare with other models

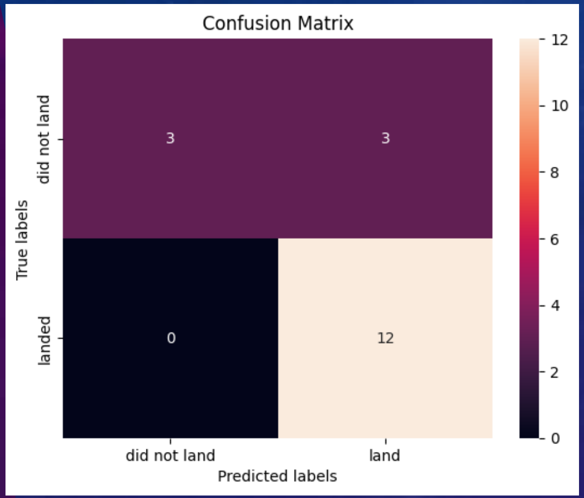
Model	Train Accuracy	Test Accuracy
Logistic Regression	0.846429	0.833333
SVM	0.848214	0.833333
DECISION TREE	0.862500	0.833333
KNN	0.848214	0.833333

RESULTS

Predictive Analytics Results

Confusion Matrics of Models

- All the models(Logistic Regression, SVM, KNN, Decision Tree) Have the same Precision and Recall



Model	Accuracy	Precision	Recall
Logistic Regression	0.833333	0.8	1.0
SVM	0.833333	0.8	1.0
DECISION TREE	0.833333	0.8	1.0
KNN	0.833333	0.8	1.0

RESULTS

Predictive Analytics Results

Efficiency of Models

- Among all the models, logistic regression shows the shortest training and prediction time, while the decision tree takes the longest

Model	Train Time	Predict Time
Logistic Regression	0.768976	0.002092
SVM	9.661876	0.003815
DECISION TREE	33.675565	0.001080
KNN	5.604476	0.007052

CONCLUSION

- KSC LC-39A: Holds the highest success rate among launch sites.
- Orbits: ES-L1, GEO, HEO, and SSO orbits have a 100% success rate.
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate.
- Launch Success: Increases over time.
- Proximity to the Equator: Launch sites closer to the equator can more effectively utilize the Earth's rotational speed, thereby reducing the energy and fuel required for launch vehicles.
- Close to the Coastline: Launch sites near the coast mean that rockets can be launched over the ocean, reducing the risk of rocket launches to densely populated areas. In the event of anomalies, rockets can safely fall into the sea rather than on densely populated land.

CONCLUSION

- Away from Densely Populated Areas: To ensure public safety, launch sites are typically located in remote areas, far from large cities and densely populated areas. This minimizes the potential impact of launch activities on public safety and the environment.
- Proximity to Infrastructure: Being close to railways and highways means that heavy equipment, rocket components, and other materials can be more easily transported to the launch sites.
- Model Performance: Among all observed models, the decision tree has the highest training accuracy, but the same test accuracy as other models. This might be a sign of overfitting. Furthermore, under the same metrics against the confusion matrix, logistic regression has the shortest training and test times compared to other models, hence showing better performance.