IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

T Hillier

27th Sep. 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection with SpaceX API and Web Scraping
  - Exploratory Data Analysis (EDA): Data Wrangling and Visualisation
  - EDA with SQL
  - Folium Interactive Map
  - Plotly Dash Dashboards
  - Predictive Analysis
- Summary of all results
  - Valuable data was collected from public sources
  - Identified features which best predicted launch success
  - Established best machine learning model to use all data to predict launch outcomes

Introduction

- Objective: evaluate the viability of new SpaceY company to compete with existing company, SpaceX
- Questions to be answered:
  - What is the best method to compute the cost of launches, given predictions of successful landings of the (reusable) first stage of the rockets?
  - What is the optimal launch site?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - SpaceX API and Web Scraping
- Perform data wrangling
  - Unnecessary columns were removed
  - Landing outcome label was added based on outcome of data after analyzing features
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Data were normalized, split into training and test sets, and evaluated using 4 different classification models – using the accuracy of each to assess their performance

Data Collection

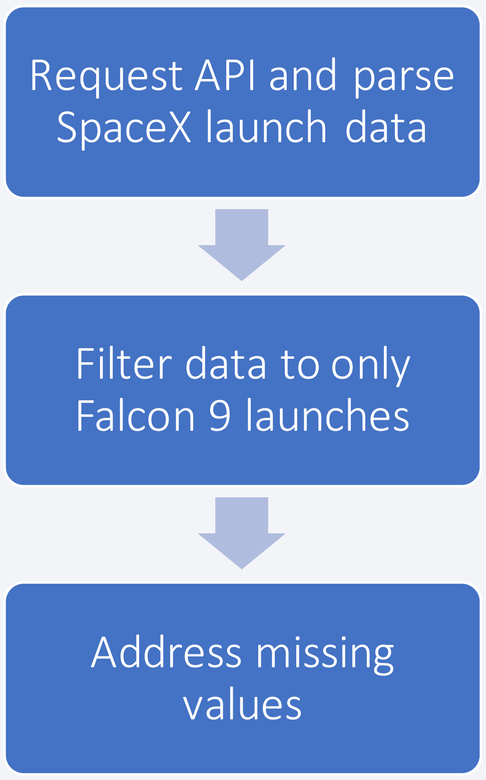
- Data sets collected from:
  - SpaceX API: https://api.spacexdata.com/v4/rockets/
  - Wikipedia: https://en.wikipedia.org/wiki/List_of_Falcon_9_and_Falcon_Heavy_launches

- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Data obtained from public SpaceX REST API

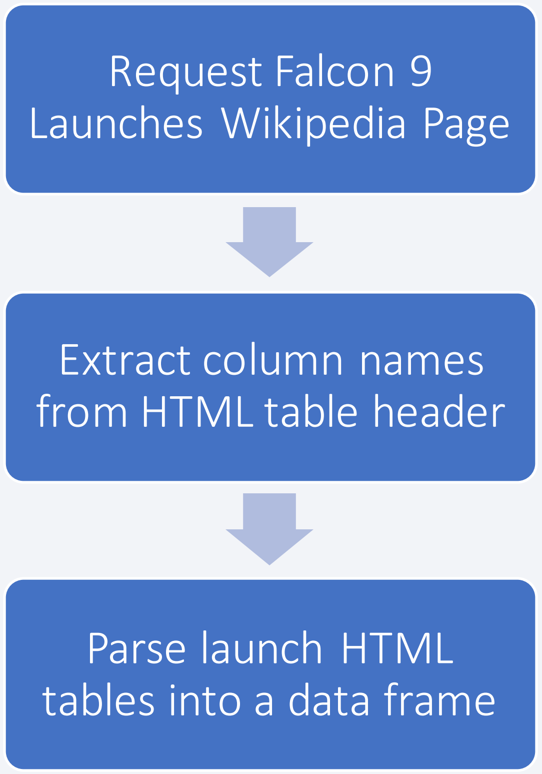
- GitHub URL of the completed SpaceX API calls notebook: https://github.com/tom-hillier/DS-Capstone/blob/master/SpaceX%20Data%20Collection.ipynb



Data Collection - Scraping

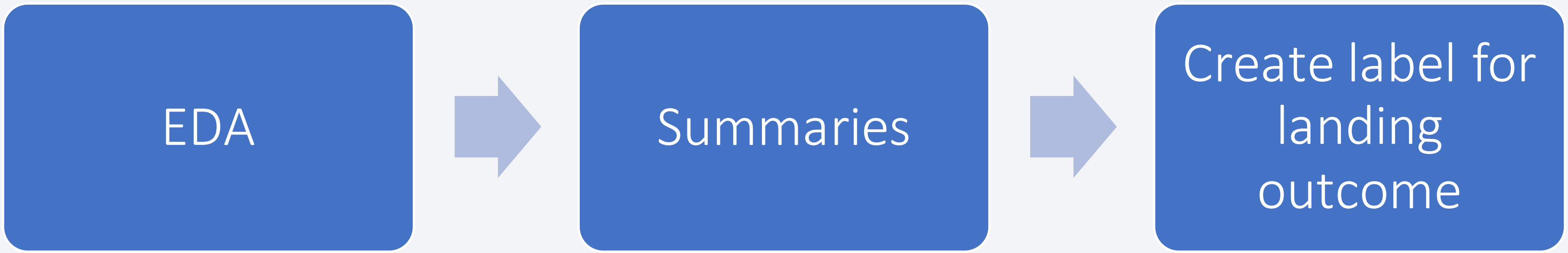
- SpaceX launch data also obtained from Wikipedia via web scraping

- GitHub URL of the completed web scraping notebook: https://github.com/tom-hillier/DS-Capstone/blob/master/SpaceX%20Web%20Scraping.ipynb



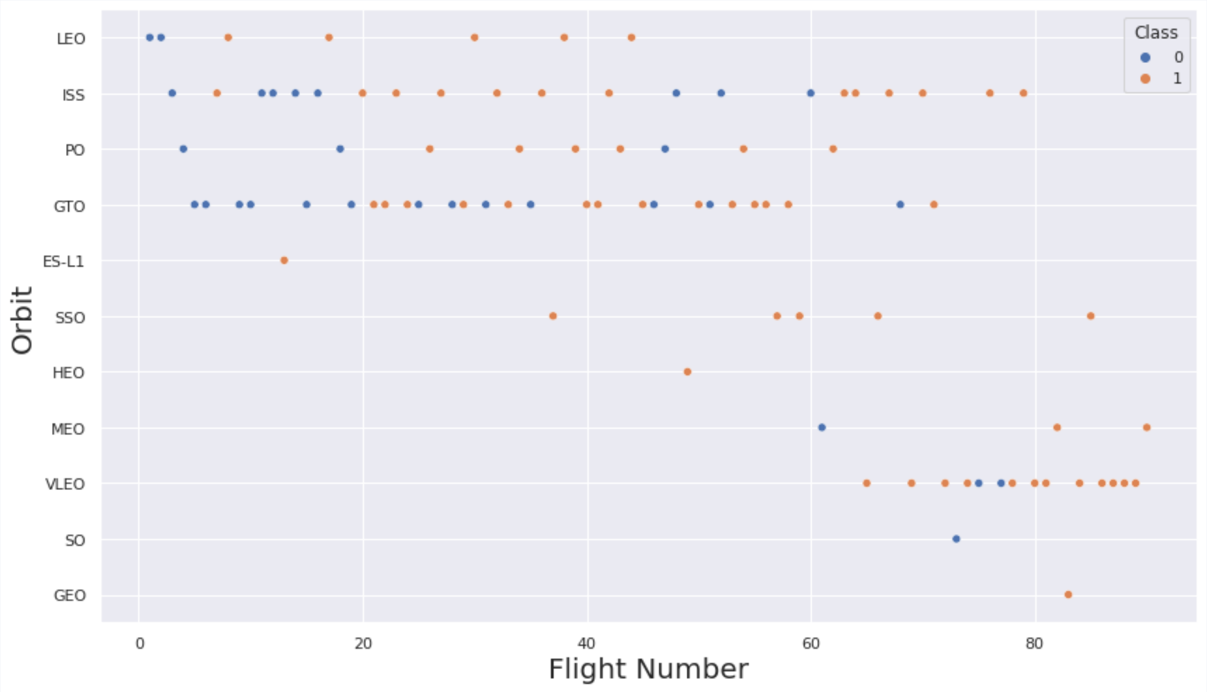
Data Wrangling

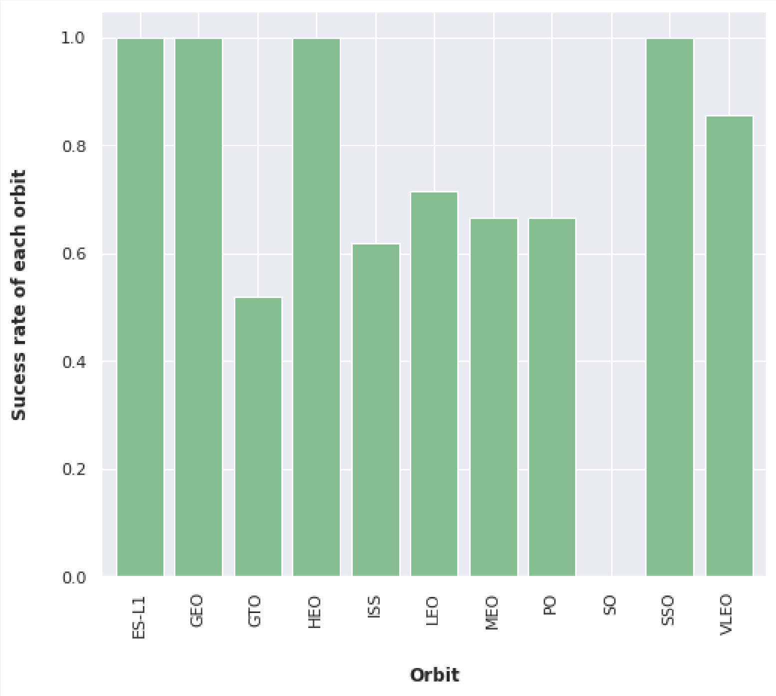
- Exploratory Data Analysis (EDA) performed on dataset
- Summaries of launches per site, occurrences of each orbit, and occurrences of mission outcome per orbit were created
- Landing outcome label created from outcome column
- GitHub URL of data wrangling related notebook: https://github.com/tom-hillier/DS-Capstone/blob/master/SpaceX%20Data%20Wrangling.ipynb



EDA with Data Visualization

- Scatter plots and bar plots used to visualize relationships in data:
  - Payload Mass, Launch Site, Orbit, Payload, Flight Number
- GitHub URL of EDA with data visualization notebook: https://github.com/tom-hillier/DS-Capstone/blob/master/EDA%20with%20Visualization.ipynb





EDA with SQL

- SQL queries performed on dataset:
  - Names of unique launch sites in the space mission
  - Top 5 launch sites with names beginning with 'CCA'
  - Total payload mass carried by boosters launched by NASA (CRS)
  - Average payload mass carried by booster version F9 v1.1
  - Data of first successful landing outcome in group pad
  - Names of the boosters which have success in drone ship and have payload mass between 4000 and 6000 kg
  - Total number of successful and failure mission outcomes
  - Names of the booster versions which carried the maximum payload mass
  - Failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015
  - Rank of the count of landing outcomes between 2010-06-04 and 2017-03-20
- GitHub URL of EDA with SQL notebook: https://github.com/tom-hillier/DS-Capstone/blob/master/EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Map objects added to Folium Map:
  - Markers to indicate points such as launch sites
  - Circles highlight areas around specific coordinates, e.g. NASA Johnson Space Center
  - Grouping of points in a cluster to display multiple and different information for similar coordinates
  - Lines used to indicate the distances between two coordinates
- These map objects allow the spatial distribution of the data to be better understood and interpreted.
- GitHub URL of interactive map with Folium map: https://github.com/tom-hillier/DS-Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

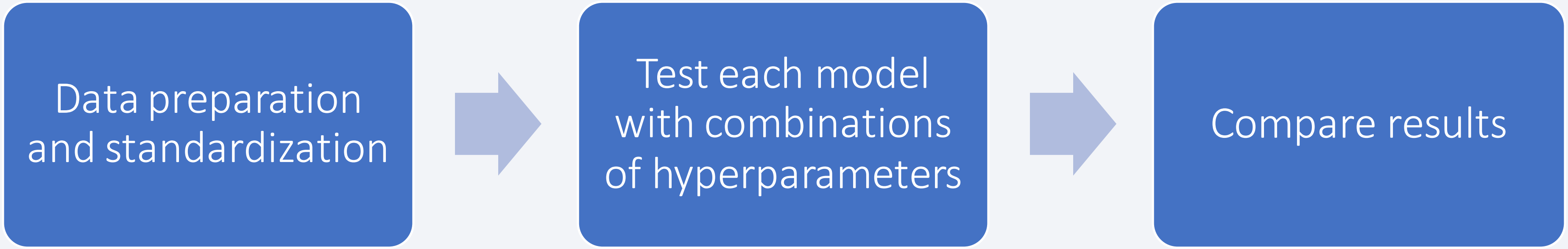
Build a Dashboard with Plotly Dash

- Dashboard has dropdown, pie chart, range slider and scatter plot components
  - Dropdown: allows user to choose launch site
  - Pie chart: shows total success/failure for selected launch site
  - Range slider: allows user to select payload mass in fixed range
  - Scatter plot: displays relationship between success and playload mass

- GitHub URL of Plotly Dash lab: https://github.com/tom-hillier/DS-Capstone/blob/master/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Predictive Analysis (Classification)

- Four classification models compared: logistic regression, support vector machine, decision tree and k-nearest neighbors
- GitHub URL of predictive analysis lab: https://github.com/tom-hillier/DS-Capstone/blob/master/Machine%20Learning%20Prediction.ipynb

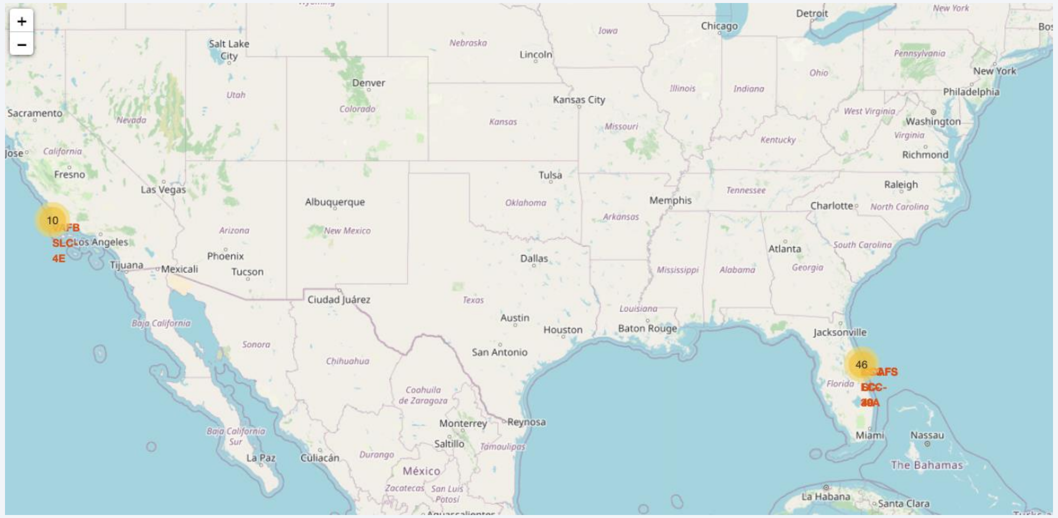


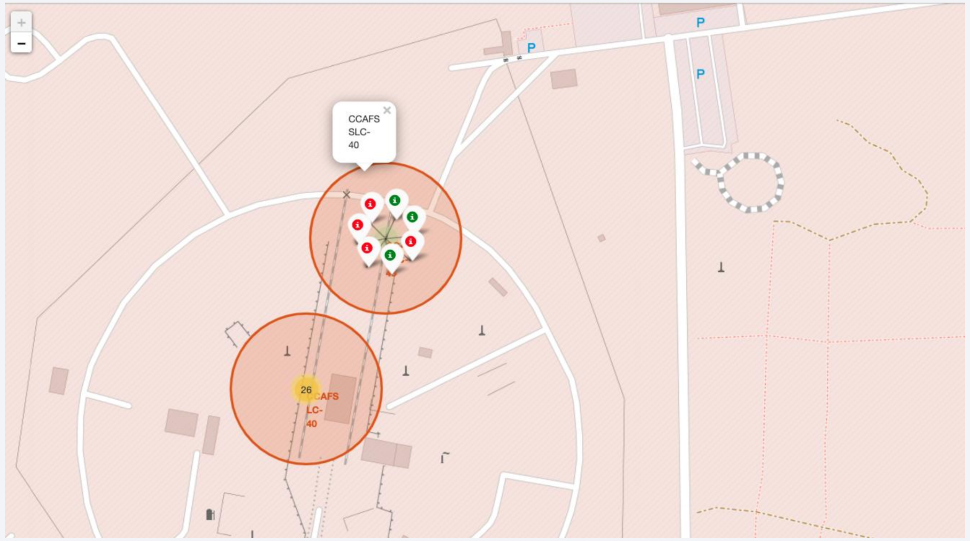
Results

- Exploratory data analysis results
  - SpaceX uses 4 launch sites
  - First launches done to SpaceX and NASA
  - Average payload of F9 v1.1 booster is 2,928 kg
  - First successful landing in 2015
  - Many Falcon 9 booster versions successful at landing in drone ships having payload above average
  - Nearly 100% of mission outcomes successful
  - 2 booster versions failed at landing in drone ships in 2015: F9 v1.1 B1012 and F9 v1.1 B1015
  - Number of landing outcomes became better over time

Results

- Interactive analytics demo in screenshots
  - Launch sites are located near the sea
  - Most launches are on the east coast, specifically Florida

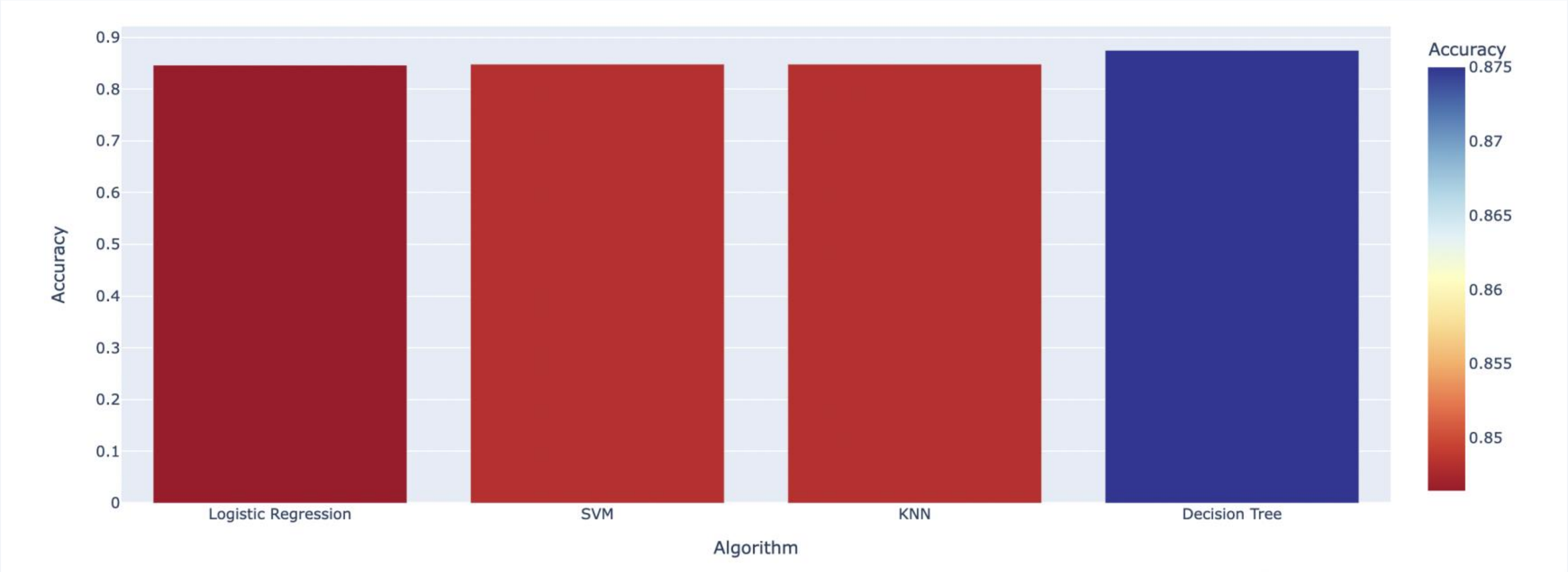




Results

- Predictive analysis results
  - Decision Tree Classifier is the best model to predict successful landings, with an accuracy of 87.5%

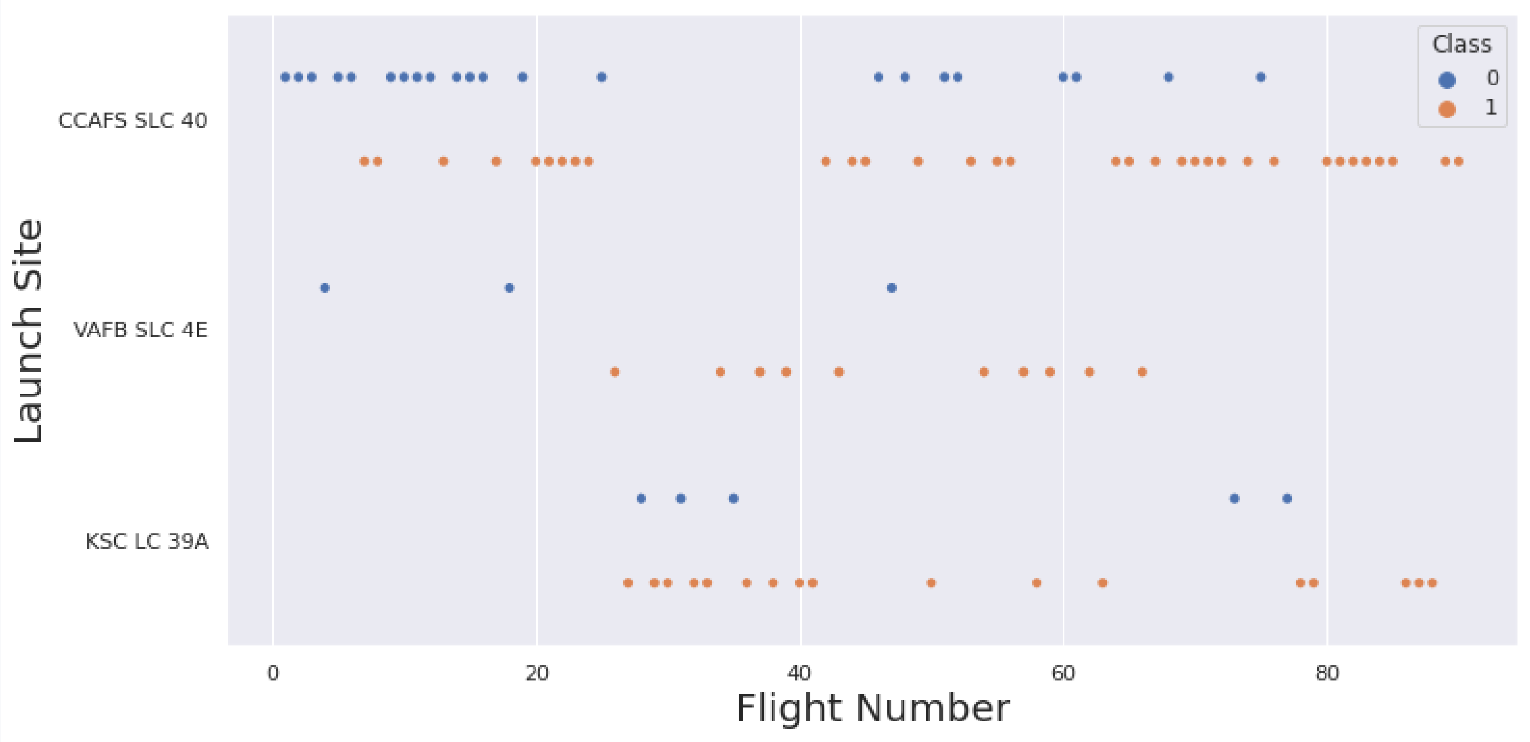
Algorithm	Accuracy
Logistic Regression	84.5%
SVM	84.8%
KNN	84.8%
Decision Tree	87.5%



Section 2

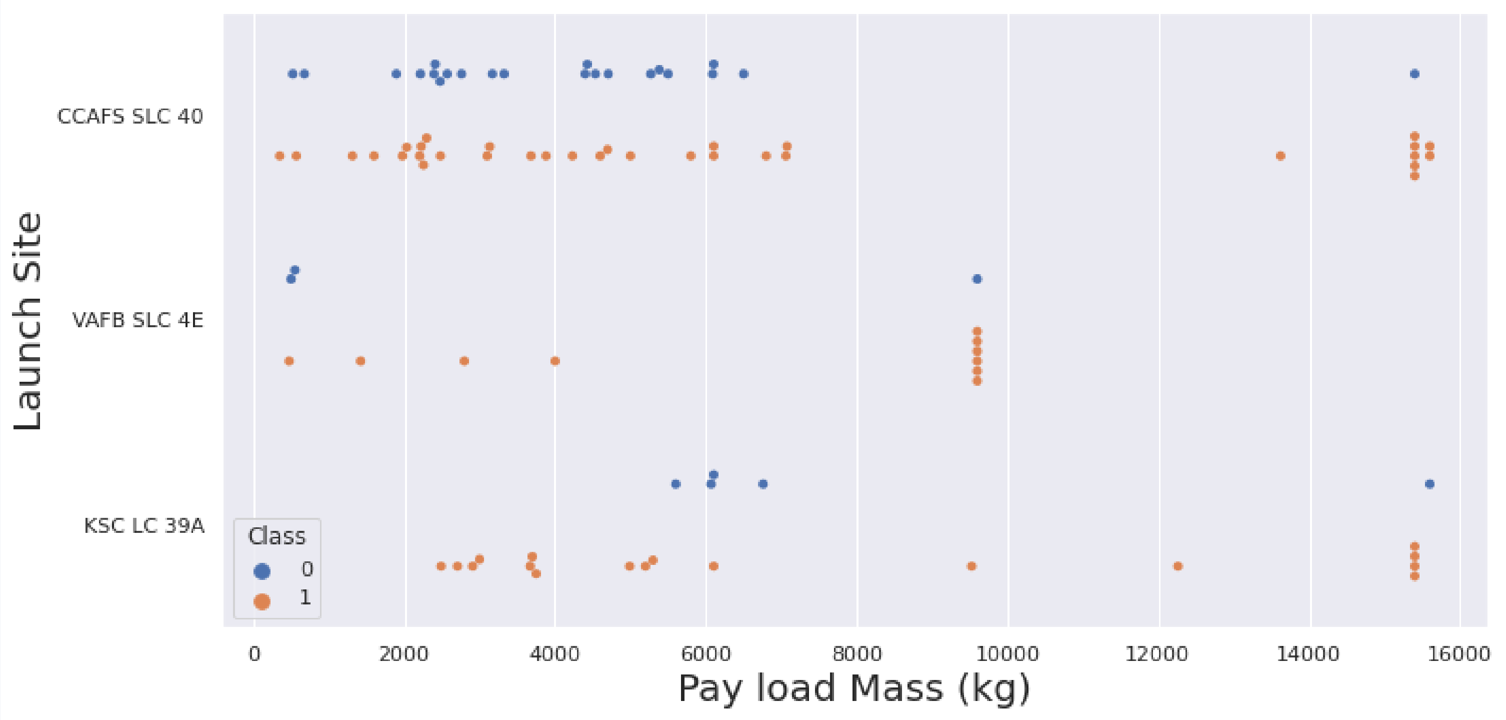
Insights drawn from EDA

Flight Number vs. Launch Site



- The general trend is that launches become more successful as the flight number increases: more Class 1 (successful, orange) than Class 0 (failure, blue).
- The most used launch site is CCAF5 SLC 40.

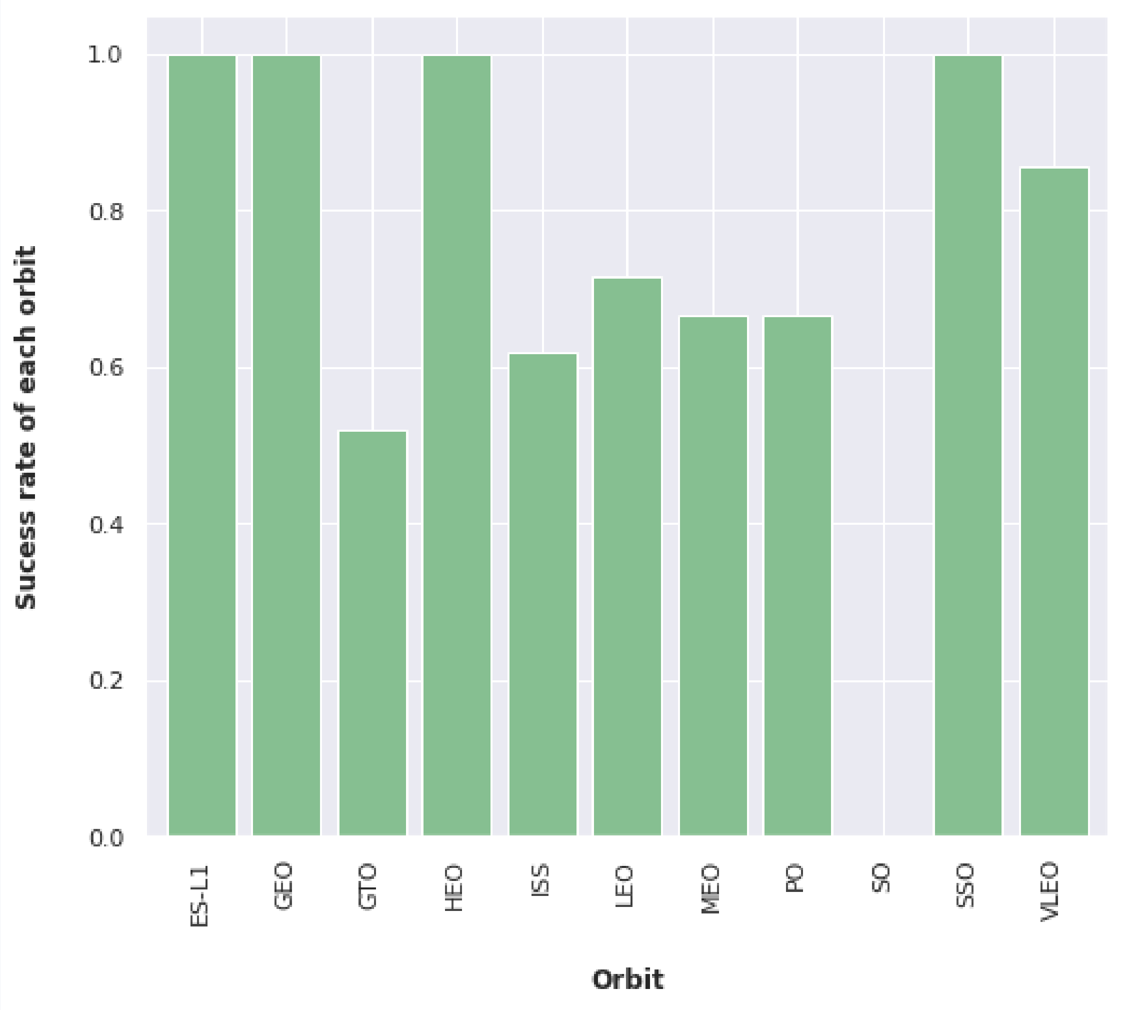
Payload vs. Launch Site



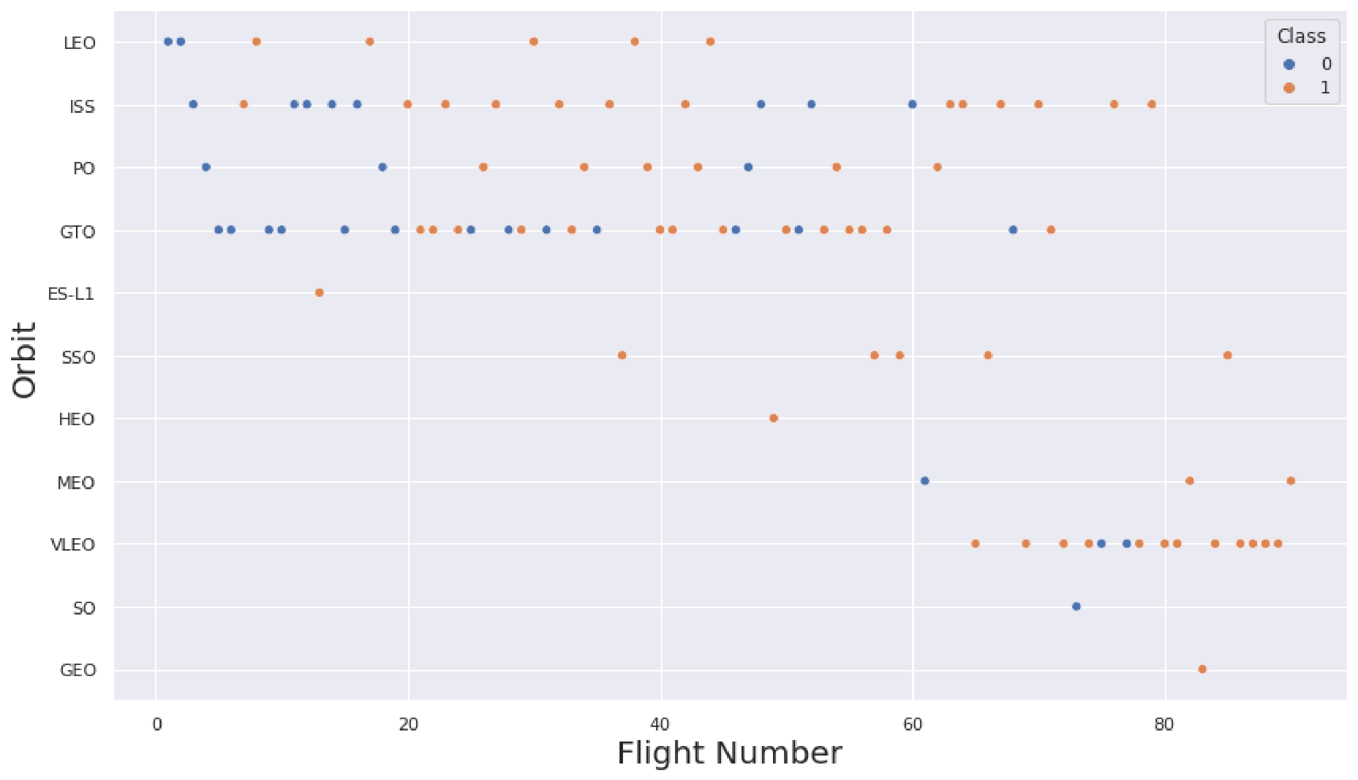
- Payloads less than 8000 kg have a fairly uniform distribution of successful/failed launches.
- Payloads over 8000 kg generally have a high success rate.

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have the best success rates
- SO has the worst, followed by GTO

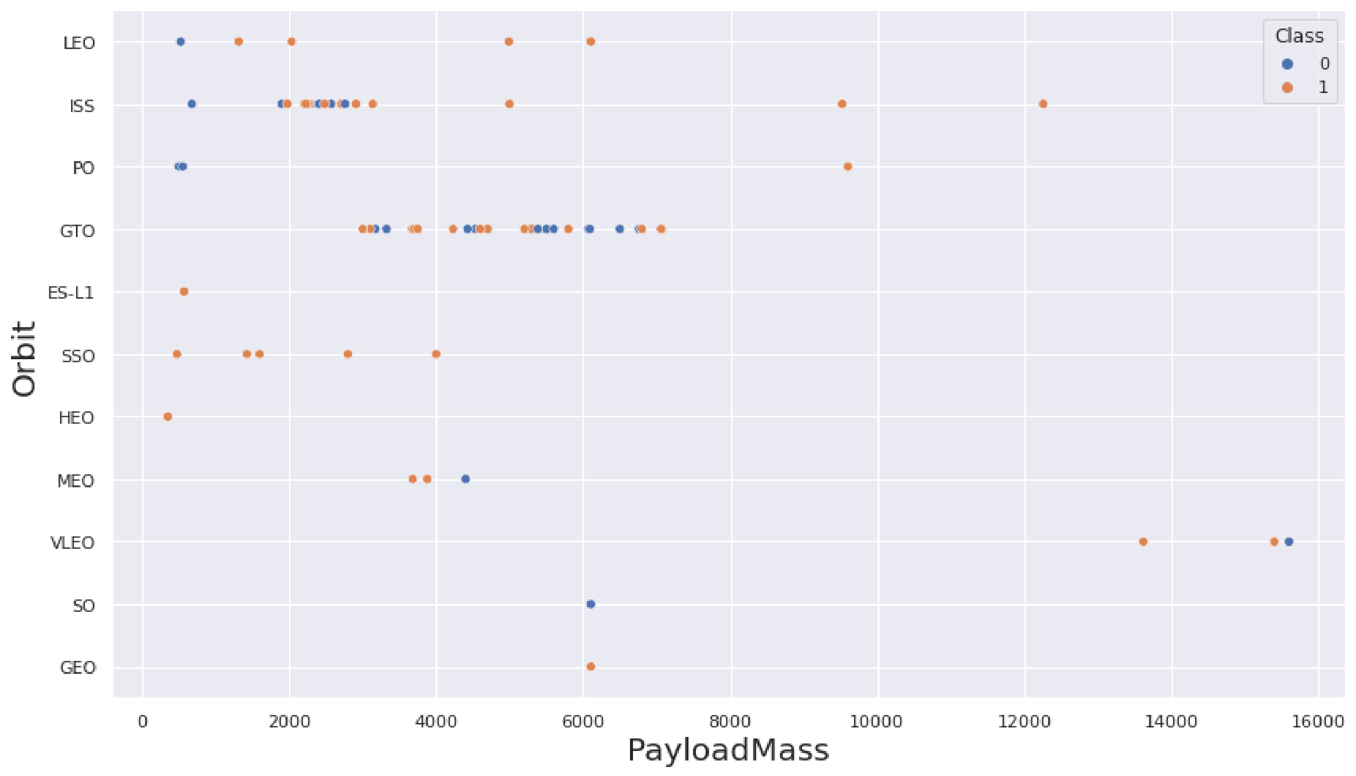


Flight Number vs. Orbit Type



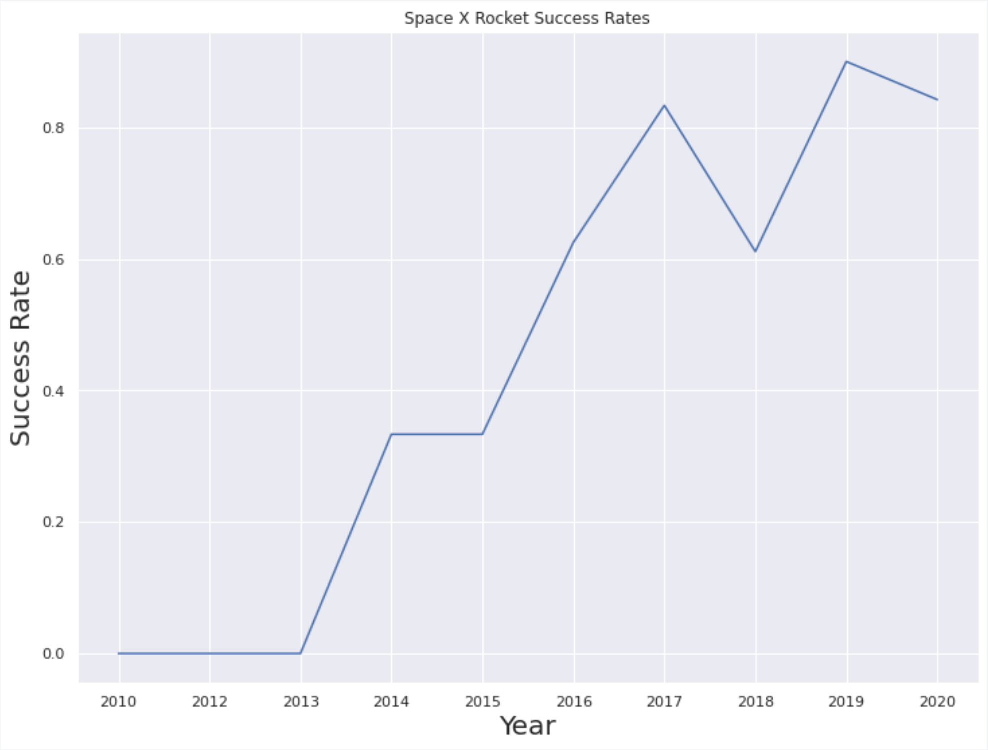
- In general, success rate improves over time for all orbits.
- However, there seems to be no correlation for the GTO orbit.

Payload vs. Orbit Type



- For heavier payloads, there is a greater landing success rate for Polar, LEO and ISS.
- For GTO, there is no discernable correlation between payload mass and orbit.

Launch Success Yearly Trend



- In general, the average success rate has steadily been increasing from 2013 to 2020.

All Launch Site Names

Finding the names of the unique launch sites

Query:

```
%sql SELECT DISTINCT "LAUNCH_SITE" FROM SPACEXDATASET
```

Result:

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

Use of DISTINCT in the SQL query ignores any duplicated in the LAUNCH_SITE column.

Query returns 4 distinct launch sites.

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`

Query:

```
%sql SELECT * FROM SPACEXDATASET WHERE "LAUNCH_SITE" LIKE '%CCA%' LIMIT 5
```

Result:

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

- Use of WHERE followed by LIKE filters launch sites containing CCA.
- Use of % within search string allows CCA to be embedded within a longer string.
- LIMIT 5 only shows 5 records.

Total Payload Mass

Query:

```
%sql SELECT SUM("PAYLOAD_MASS__KG_") FROM SPACEXDATASET WHERE "CUSTOMER" = 'NASA (CRS)'
```

Result:

1
45596

Use of SUM to total the PAYLOAD_MASS_KG_ for the case where the customer is NASA (CRS)

Average Payload Mass by F9 v1.1

Query:

```
%sql SELECT AVG("PAYLOAD_MASS__KG_") FROM SPACEXDATASET WHERE "BOOSTER_VERSION" LIKE '%F9 v1.1%'
```

Result:

1
2534

Use of AVG to average the PAYLOAD_MASS_KG_ for the case where the booster version contains the substring F9 v1.1

First Successful Ground Landing Date

Query:

```
%sql SELECT MIN("DATE") FROM SPACEXDATASET WHERE "LANDING__OUTCOME" LIKE '%Success%'
```

Result:

1
2015-12-22

This query finds the oldest successful landing, selected by the MIN function.

The WHERE clause filters only successful landings.

Successful Drone Ship Landing with Payload between 4000 and 6000

Query:

```
%sql SELECT BOOSTER_VERSION FROM SPACEXDATASET WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

Result:

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

Query returns the booster versions where landing on a drone ship was successful, and the payload mass is between 4000 and 6000 kg. These conditions are filtered by the WHERE and AND clauses.

Total Number of Successful and Failure Mission Outcomes

Query:

```
%sql SELECT sum(case when MISSION_OUTCOME LIKE '%Success%' then 1 else 0 end) AS "Successful Mission", \
    sum(case when MISSION_OUTCOME LIKE '%Failure%' then 1 else 0 end) AS "Failure Mission" \
FROM SPACEXDATASET;
```

Result:

Successful Mission	Failure Mission
100	1

Query finds all successful missions, and sums them to the "Successful Mission" column; then it does the same for failures and sums them to the "Failure Mission" column.

Boosters Carried Maximum Payload

Query:

```
%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEXDATASET \
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

Result:

Booster Versions which carried the Maximum Payload Mass
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

The subquery filters the data by only returning the heaviest payload masses – using the MAX function.

The main query uses the results of the subquery and returns unique booster versions with the heaviest payload masses.

2015 Launch Records

Query:

```
%sql SELECT {fn MONTHNAME(DATE)} as "Month", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXDATASET WHERE year(DATE) = '2015' AND \
LANDING__OUTCOME = 'Failure (drone ship)';
```

Result:

Month	booster_version	launch_site
January	F9 v1.1 B1012	CCAFS LC-40
April	F9 v1.1 B1015	CCAFS LC-40

Query returns the month, booster version and launch site where landing was unsuccessful; for the year 2015.

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

Query:

```
%sql SELECT LANDING__OUTCOME as "Landing Outcome", COUNT(LANDING__OUTCOME) AS "Total Count" FROM SPACEXDATASET \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC ;
```

Result:

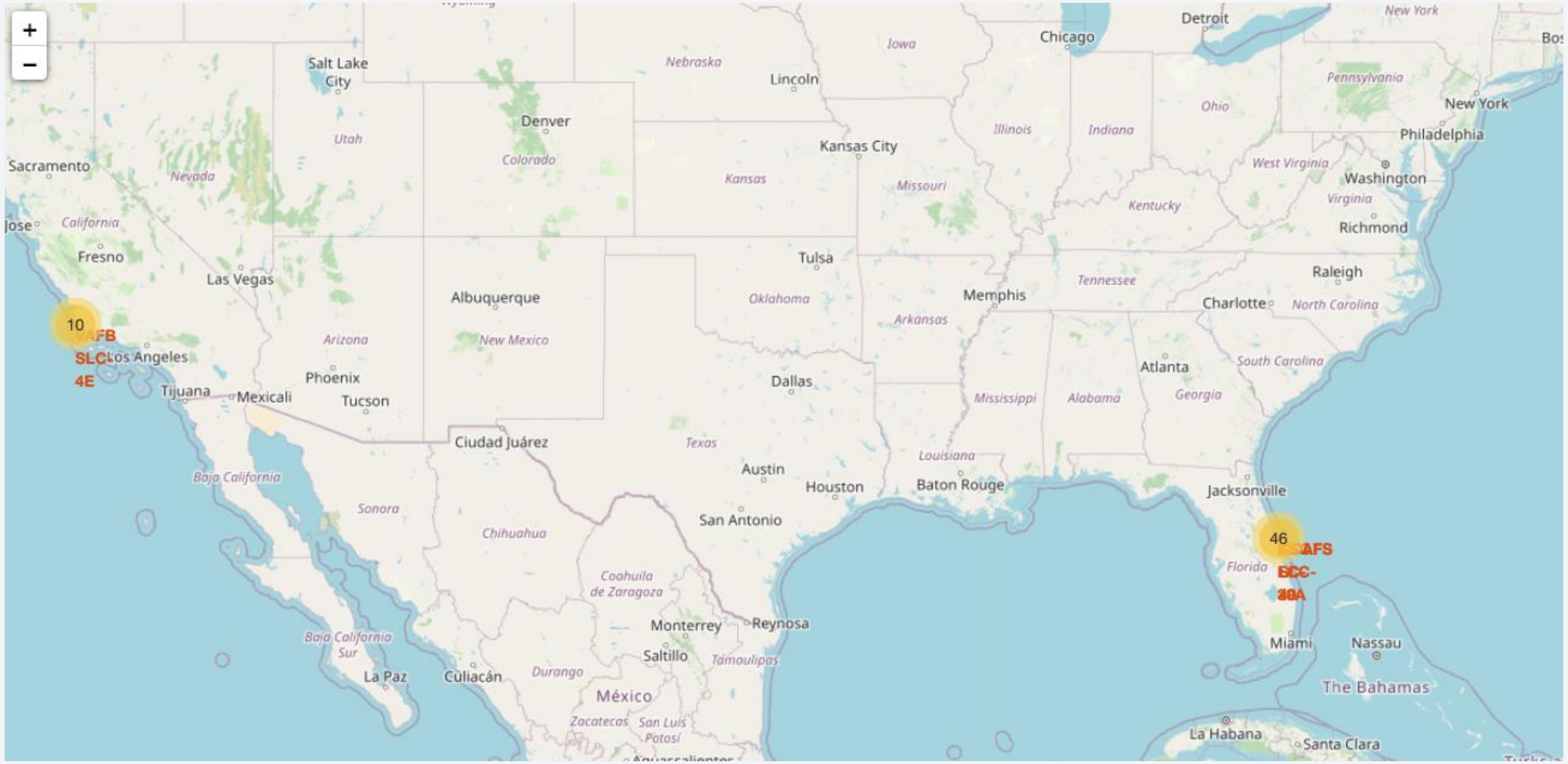
Landing Outcome	Total Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Query returns count of landing outcomes between the dates 2010-06-04 and 2017-03-20, and displays them in descending order. GROUP_BY groups results by landing outcome, and ORDER BY COUNT DESC displays the results in descending order.

Section 3

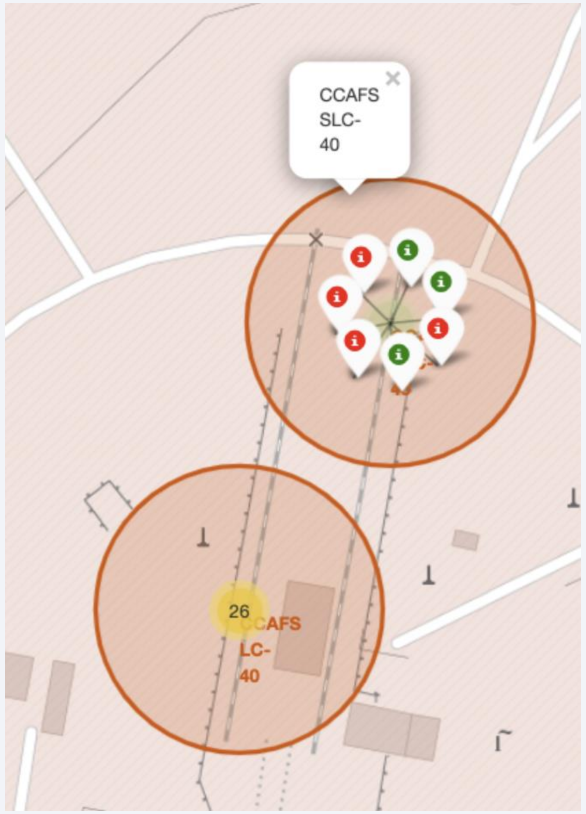
Launch Sites Proximities Analysis

Folium Map: Overview of All Launch Sites



SpaceX launch sites are located on the coast of the US, where most are concentrated on the east coast, in Florida.

Folium Map: Colored Markers

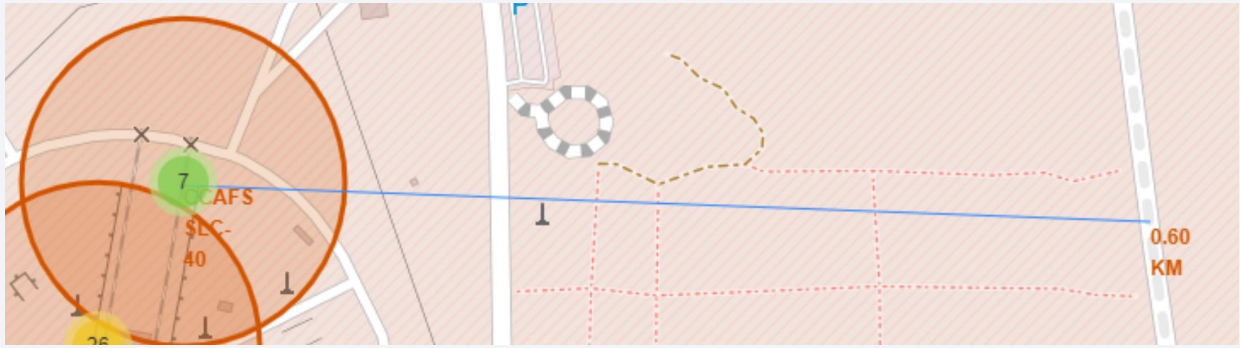


Zoom in of launch site CCAFS SLC-40, showing the sites of successful (green) and unsuccessful (red) launches.

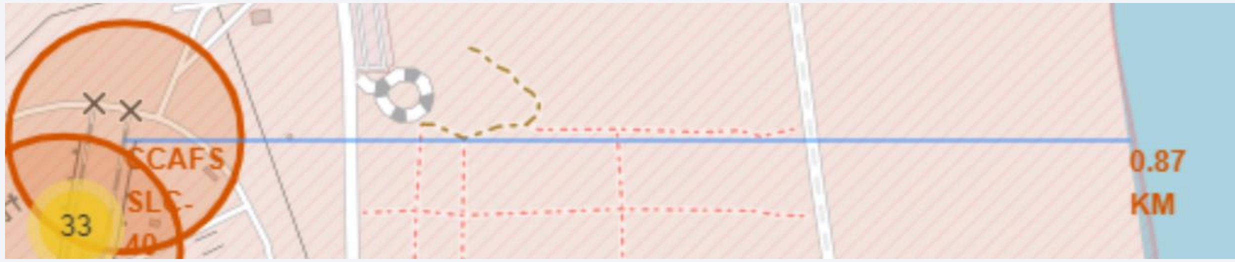
Folium Map: CCAFS SLC-40 Distances

Proximity to railway: 1.29 km

Proximity to road: 0.6 km



Proximity to coastline: 0.87 km

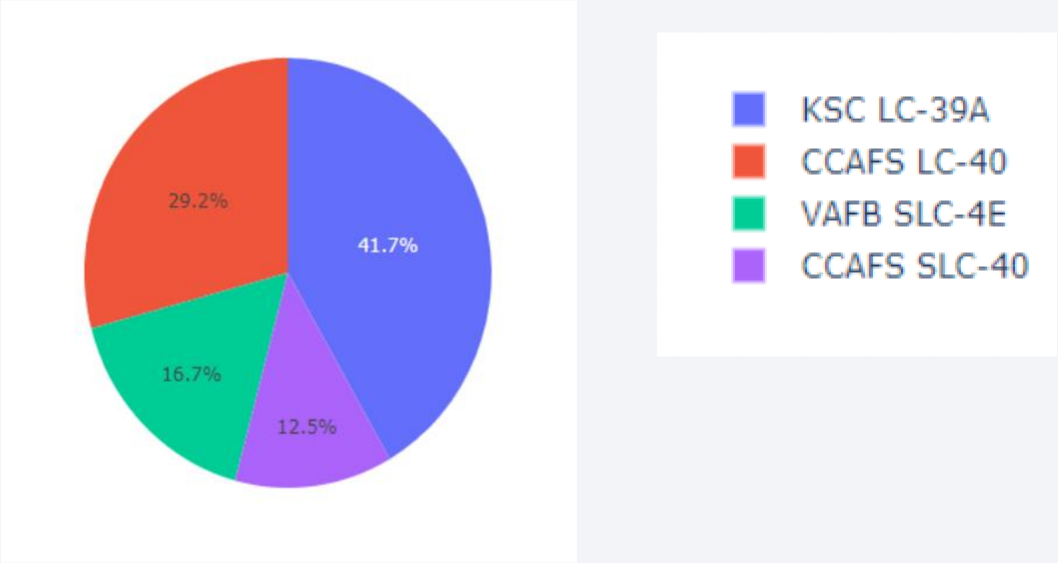


The launch site CCAFS SLC-40 is fairly close to the coastline, as noted previously. It is also close to a road and railway, as would be expected. The launch site is a large distance of approximately 23 km from the nearest city, which would also be expected.

Section 4

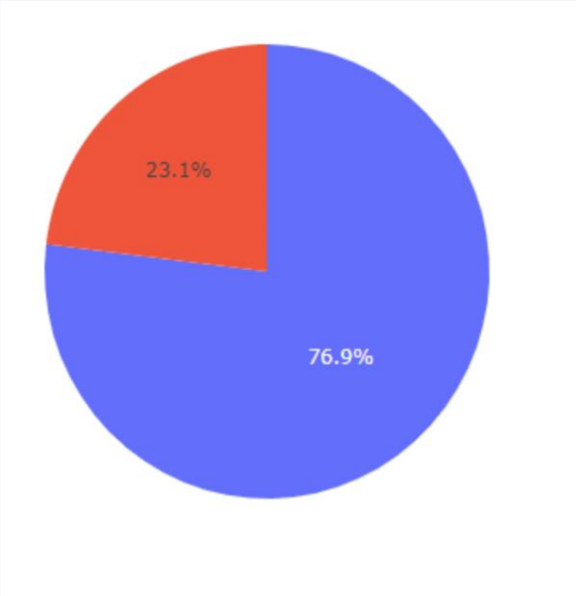
Build a Dashboard with Plotly Dash

Dashboard: Launch Site Success



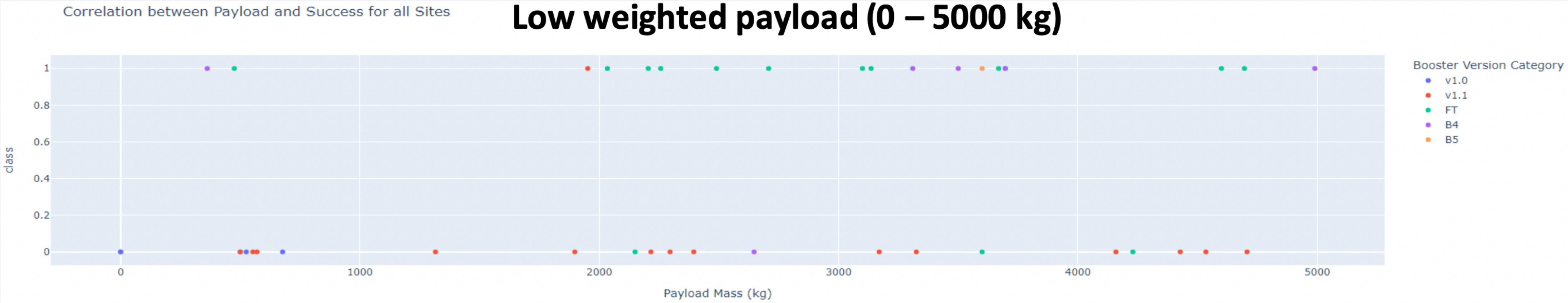
The pie chart shows the most successful launch site to be KSC LC-39A

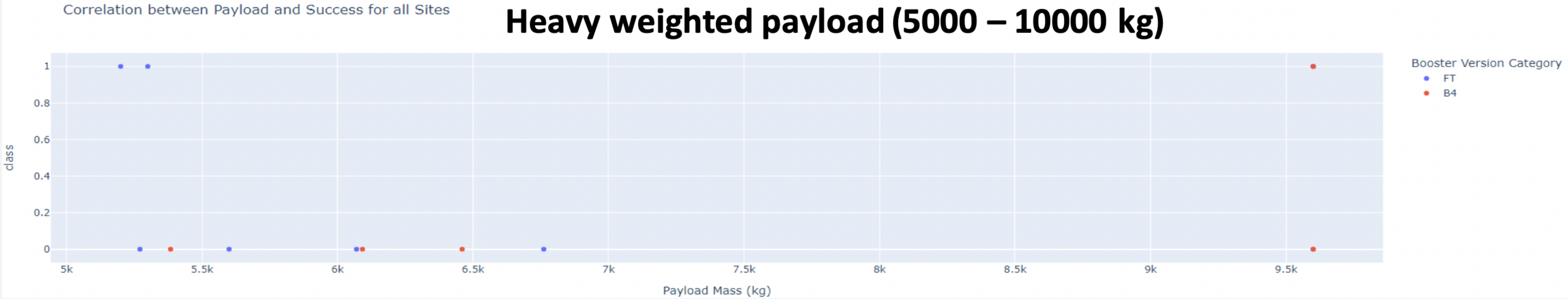
Dashboard: Launch Results for KSC LC-39A



KSC LC-39A has a launch success rate of about 77%.

Dashboard: Payload vs. Launch Outcome





Payload vs. Launch Outcome scatter plots for all site, with low weighted (0 – 5000 kg) and heavy weighted (5000 – 10000 kg) payloads selected on the range slider.

Low weighted payloads generally have a higher success rate than heavy weighted payloads

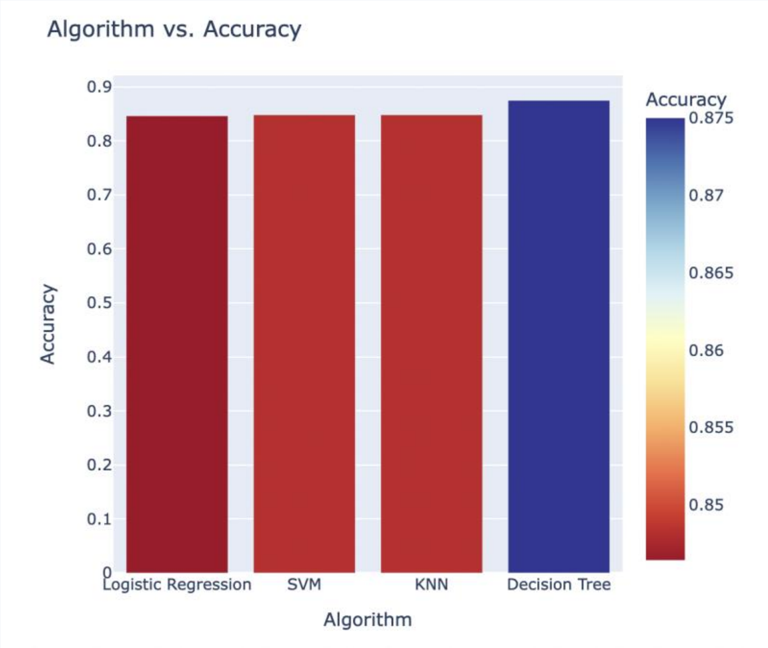
Section 5

Predictive Analysis (Classification)

Classification Accuracy

Model accuracies for all algorithms, presented as a bar chart. The Decision Tree algorithm performs the best, with a classification accuracy of 87.5%.

	Algorithm	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	KNN	0.848214
3	Decision Tree	0.875000

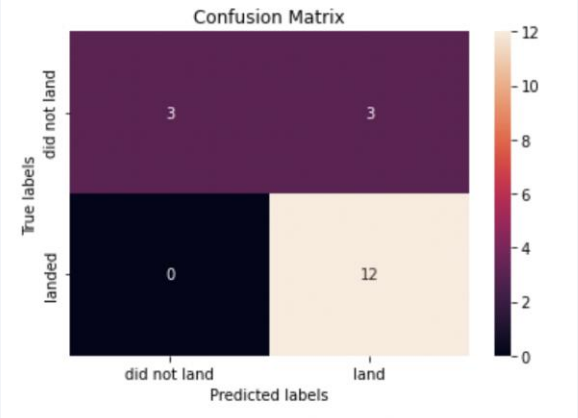


Decision Tree best parameters:

```
tuned hpyerparameters :(best parameters)  {'criterion': 'gini', 'max_depth
': 4, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 1
0, 'splitter': 'random'}
accuracy : 0.875
```

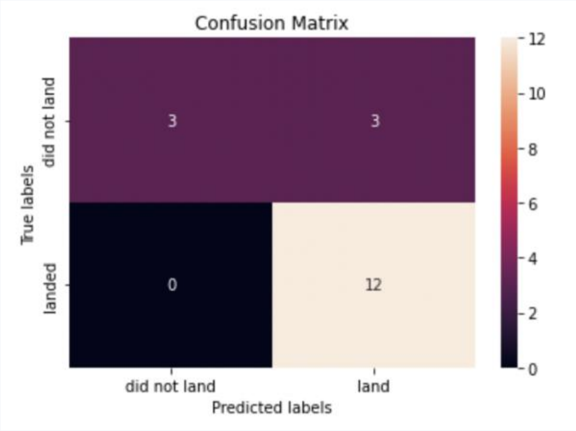
Confusion Matrix

Decision Tree

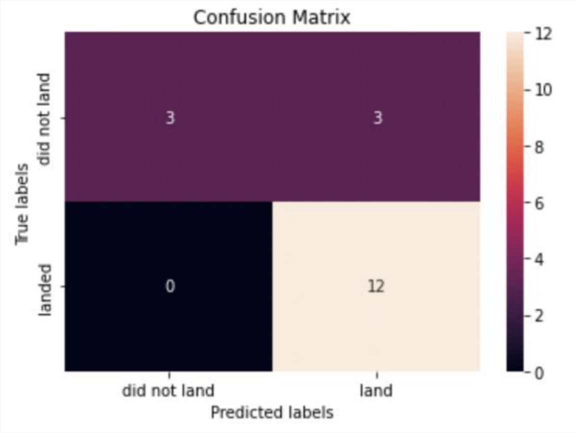


The confusion matrices for all models are identical. The main issue which should be addressed in further work is the non-zero entries in the false positives cell – the matrices should all ideally be diagonal.

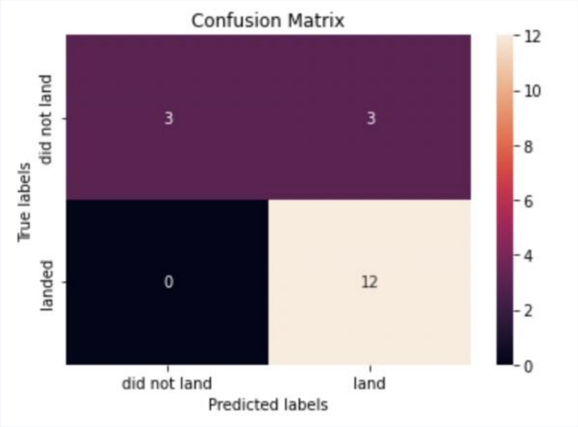
Logistic Regression



SVM



kNN



Conclusions

- Mission success can be explained by several dependent factors including the launch site, orbit type, number of previous launches, payload mass etc.
- In general, light payload masses perform better than heavy payloads.
- Orbits ES-L1, GEO, HEO and SSO have the highest success rates.
- Current data cannot explain the differences in success rates between launch sites. More data is possibly required, such as weather reports.
- To predict future launches, the Decision Tree algorithm should be used since it gave the best accuracy on the training data.

Thank you!