IBM DATA SCIENCE CAPSTONE PROJECT

LPNM

Executive Summary

Objective

This study aims to identify the factors for a successful rocket landing

Summary of Methodologies

- Collect data using SpaceX REST API and web scraping
- Data Wrangling
- Exploratory data Analysis
- Interactive Visual Analytics
- Predictive Analysis

Results

- Exploratory data Analysis:
  1. Launch success has improved over time
  2. KSC LC-39A is the landing site with the highest success rate
  3. Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
- Visualization Analytics
  1. Most launch sites are near the equator and close to the coast
- Predictive Analysis
  1. The decision tree model showed the best performance among the tested models

INTRODUCTION

- SpaceX launches Falcon 9 rockets at a cost of around $62m. This is considerably cheaper than other providers, and much of the savings are because SpaceX can land and then re-use the first stage of the rocket.

- If we can make predictions on whether the first stage will land, we can determine the cost of a launch, and use this information to assess whether or not an alternate company should bid against SpaceX for a rocket launch.

- This project aims to produce the best model to predict whether a landing will be successfull or not

Methodology

- Collect data using SpaceX Rest API and web scraping techniques
- Wrangle data by filtering the data, handling missing values and applying one hot encoding to prepare the data for analysis and modeling
- Explore data via EDA with SQL and data visualization techniques
- Visualize the data using Folium and Plotly Dash
- Build Models to predict landing outcomes using classification models.
- Tune and evaluate models to find best model

Data Collection - API

- Request data from SpaceX API
- Decode response using .json() and convert to a dataframe using .json_normalize()
- Request information about the launches from SpaceX API using custom functions
- Create dictionary from the data
- Create dataframe from the dictionary
- Filter dataframe to contain only Falcon 9 launches
- Replace missing values of Payload Mass with .mean()
- Export data to .csv file

Data Collection – Web Scraping

- Request Falcon 9 launch data from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from HTML table header
- Collect data from parsing HTML tables
- Create dictionary from the data
- Create dataframe from the dictionary
- Export data to csv file

Data Wrangling

Steps

- Perform EDA and determine the labels
- Calculate number of: launches per site, occurrence of orbit, mission outcome per orbit type
- Create binary landing outcome column (1 for successful 0 for unsuccessful
- Export data to .csv file

Landing Possible Outcomes:

- True Ocean: mission outcome had a succesful landing to a specific region of the ocean
- False Ocean: represents an unsuccessful landing to a specific region of the ocean
- True RTLS: means the mission had a successful landing on a ground pad
- False RTLS: represents an unsuccessful landing on a ground pad
- True ASDS: represents the mission had a successful landing on a drone ship
- False ASDS: Means an unsuccessful landing on a drone ship

EDA with Visualization

Charts

- Flight number vs. Payload
- Flight number vs. Launch Site
- Payload mass (kg) vs. Launch Site
- Payload mass (kg) vs. Orbit Type

Analysis

- View relationship by using scatter plots. The variables could be useful for machine learning if a relationship exists.

- Show comparisons among discrete categories with bar charts. Bar charts show the relationships among the categories and a measured value

EDA with SQL

Display

- Names of unique launch sites
- 5 records where launch site begins with 'CCA'
- Total payload mass carried by boosters launched by NASA
- Average payload mass carried by booster version F9 v1.1.

List

- Date of first succesful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4000kg but less than 6000kg
- Total number of successful and failed missions
- Names of booster versions which have carried the max payload
- Failed landing outcomes on drone ships, their booster version and launch site for the months of 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20

Map with Folium

Markers indicating launch sites

- Added blue circle at NASA Johnson Space Center's coordinates with a popup label showing its name and coordinates
- Added red circles at all launch sites coordinates with a popup label showing its names and coordinates

Colored Markers of Launch Outcomes

- Added colored markers of successful (green) and unsuccessful (red) launches at each launch site to show which have higher success rates

Distances Between a Launch Site and Proximities

- Added colored lines to show distance between launch site CCAFS SLC-40 and its proximity to the nearest coastline, railway, highway and city.

Dashboard with Plotly Dash

Dropdown List with Launch Sites

- Allow user to select all launch sites or a certain launch site

Pie Chart Showing Successful Launches

- Allow user to see successful and unsuccessful launches as a percent of the total

Slider of Payload Mass Range

- Allow user to select payload mass range

Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version

- Allow user to see the correlation between Payload and Launch Success

Predictive Analytics

- Create NumPy array for Class column
- Standardize the data with StandardScaler. Fit and trnasform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optmization
- Apply GridSearchCV on different algorithms: logistic regression, support vector machine, decision tree, K-nearest Neighbor
- Calculate accuracy on the test data using .score() for all models
- Assess the confusion matrix for all models
- Identify the best model using Jaccard_Score, F1_score and Accuracy

Results Summary

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough from anything a failed launch could damage, while still well connected to logistics infrastructure

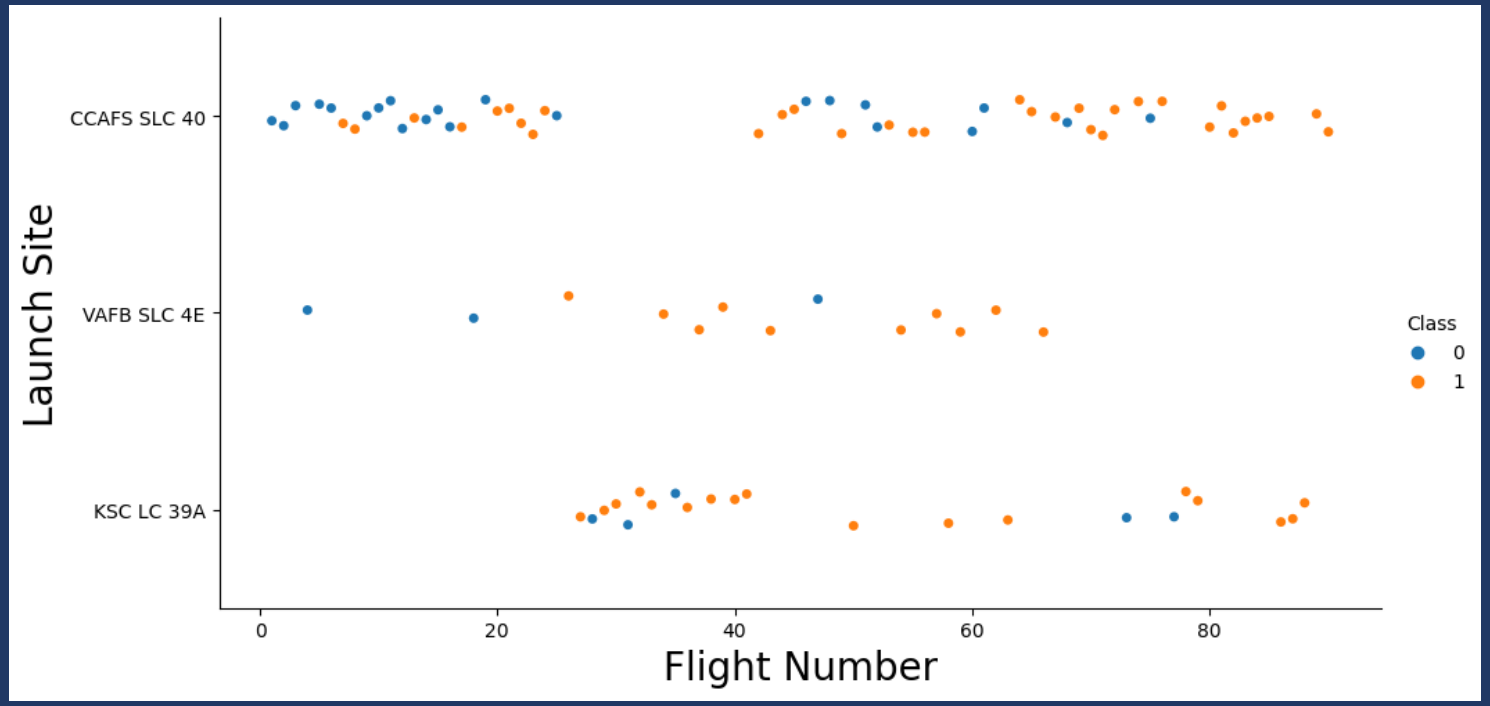
Predictive Analytics

- Decision Tree model is the best predictive model for the dataset

Flight Number vs. Launch Site

Exploratory Data Analysis

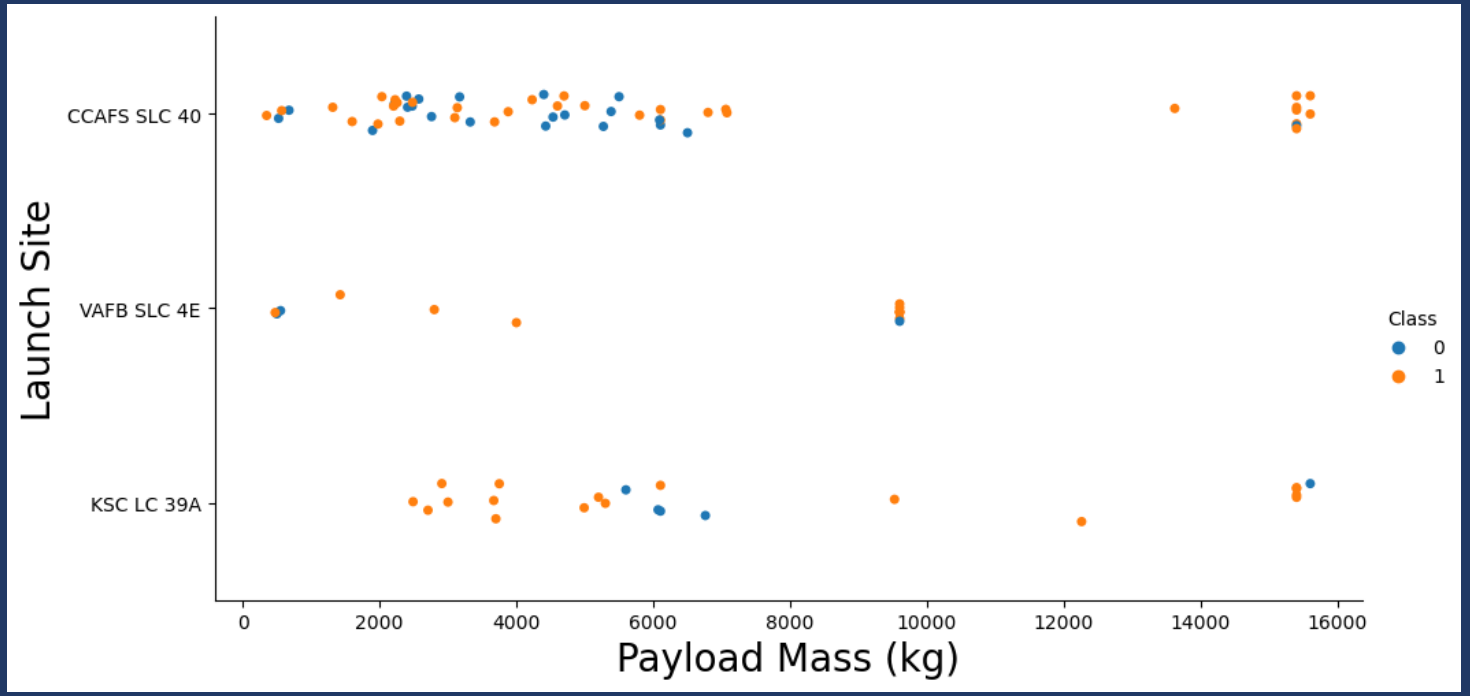
- Earlier flights had a lower success rate
- Later flights had a higher success rate
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates



Payload vs. Launch Site

Exploratory Data Analysis

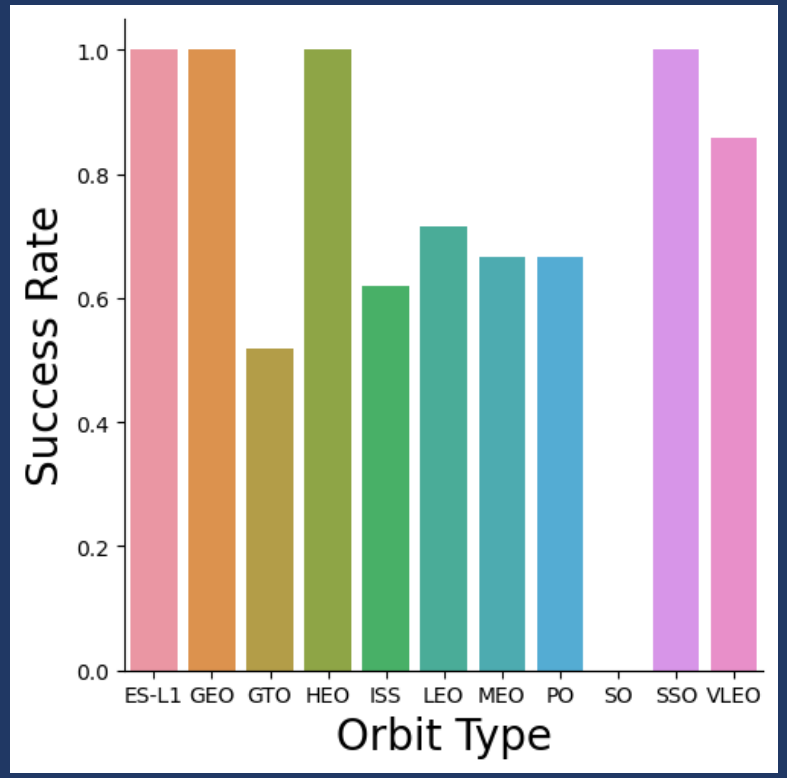
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5500kg
- VAFB SKC 4E has not launched anything greater than 10000kg



Success Rate by Orbit

Exploratory Data Analysis

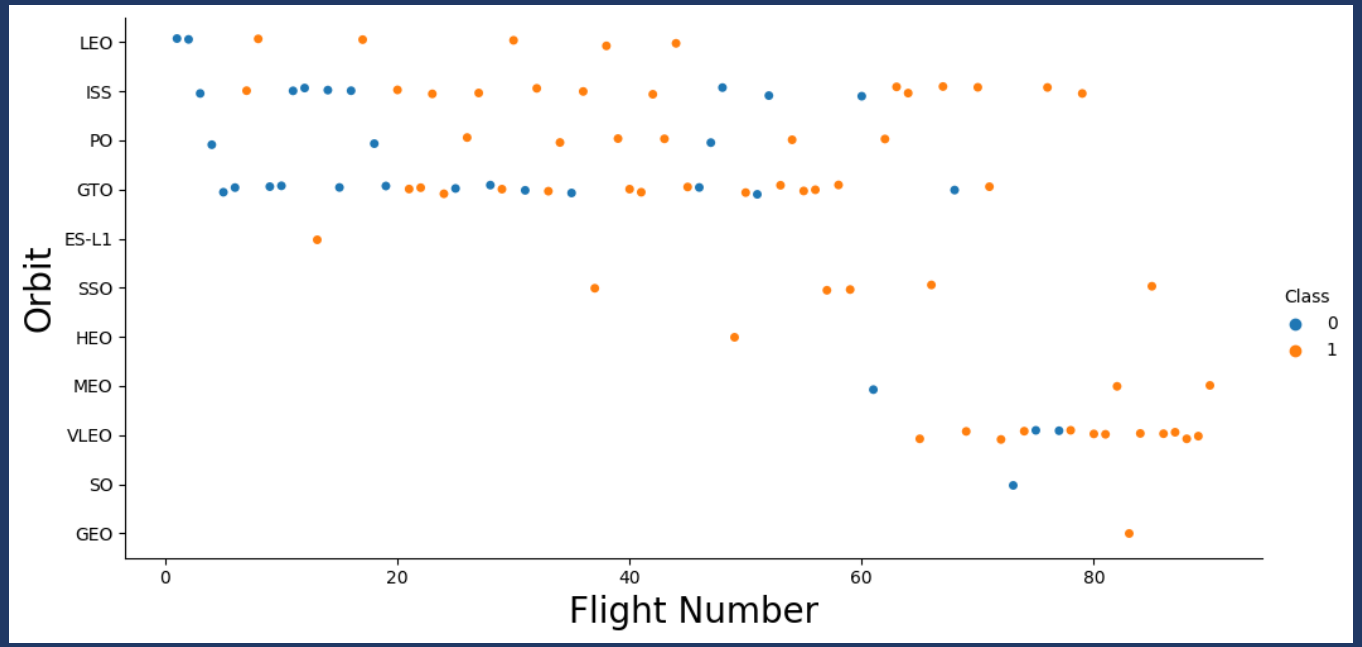
- 100% Success Rate: ES-L1, GEO, HEO, SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



Flight Number vs. Orbit

Exploratory Data Analysis

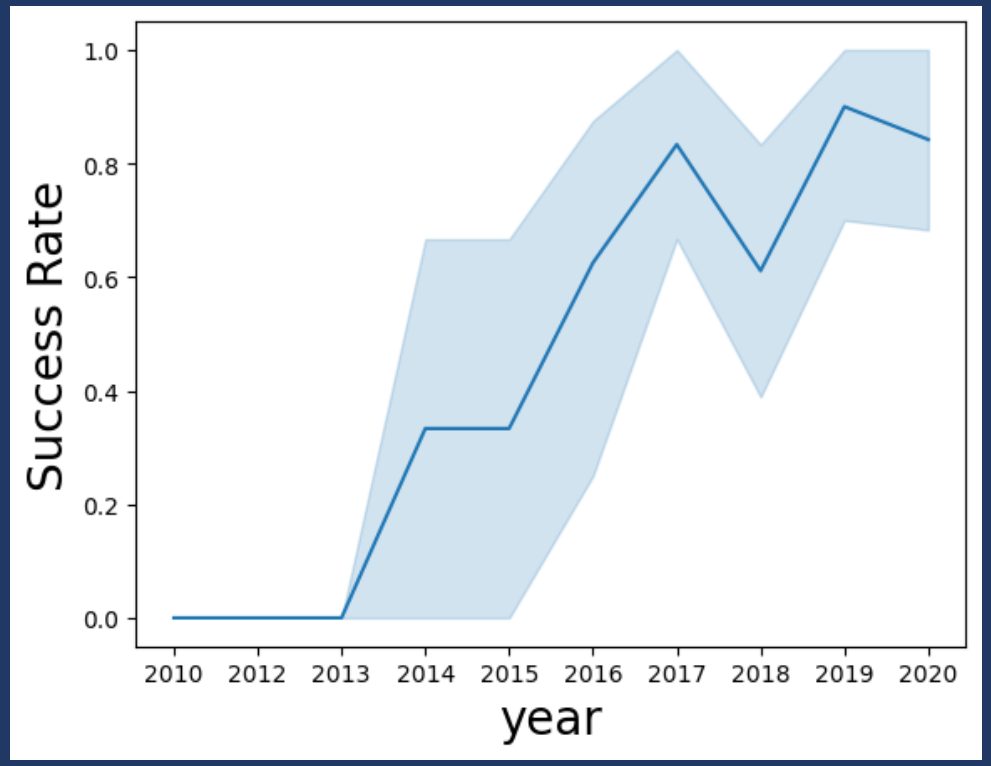
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit does not follow this trend



Launch Success over Time

Exploratory Data Analysis

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



Launch Site Information

Launch Site Names

```
%sql SELECT DISTINCT(LAUNCH_SITE) FROM SPACEXTBL;
```

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

Records with Launch Site starting with CCA

```
%sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Payload Mass

Total Payload Mass

- 45596kg carried by boosters launched by NASA(CRS)

```
%sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS FROM SPACEXTBL
    WHERE CUSTOMER = 'NASA (CRS)';
```

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

TOTAL_PAYLOAD_MASS
45596

Average Payload Mass

- 2928kg on average carried by booster version F9 v1.1

```
%sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVERAGE_PAYLOAD_MASS FROM SPACEXTBL
    WHERE BOOSTER_VERSION = 'F9 v1.1';
```

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

AVERAGE_PAYLOAD_MASS
2928.4

Landing info

1st Successful landing on ground pad

```
%sql SELECT MIN(DATE) AS FIRST_SUCCESSFUL_GROUND_LANDING FROM SPACEXTBL
    WHERE Landing_Outcome = 'Success (ground pad)';
```

FIRST_SUCCESSFUL_GROUND_LANDING
2015-12-22

Booster drone ship landing

```
%sql SELECT BOOSTER_VERSION FROM SPACEXTBL \
    WHERE (LANDING_OUTCOME = 'Success (drone ship)') AND (PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000);
```

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

Total number of Successful and failed missions

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

Boosters carrying max Payload

```
%sql SELECT DISTINCT(BOOSTER_VERSION) FROM SPACEXTBL \
    WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

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

Failed Landings on Drone Ship

During 2015

```
%%sql SELECT substr(Date,6,2) as Month, DATE,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome
FROM SPACEXTBL
where (Landing_Outcome = 'Failure (drone ship)') and substr(Date,1,4)='2015';
```

Month	Date	Booster_Version	Launch_Site	Landing_Outcome
10	2015-10-01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Successful Landings

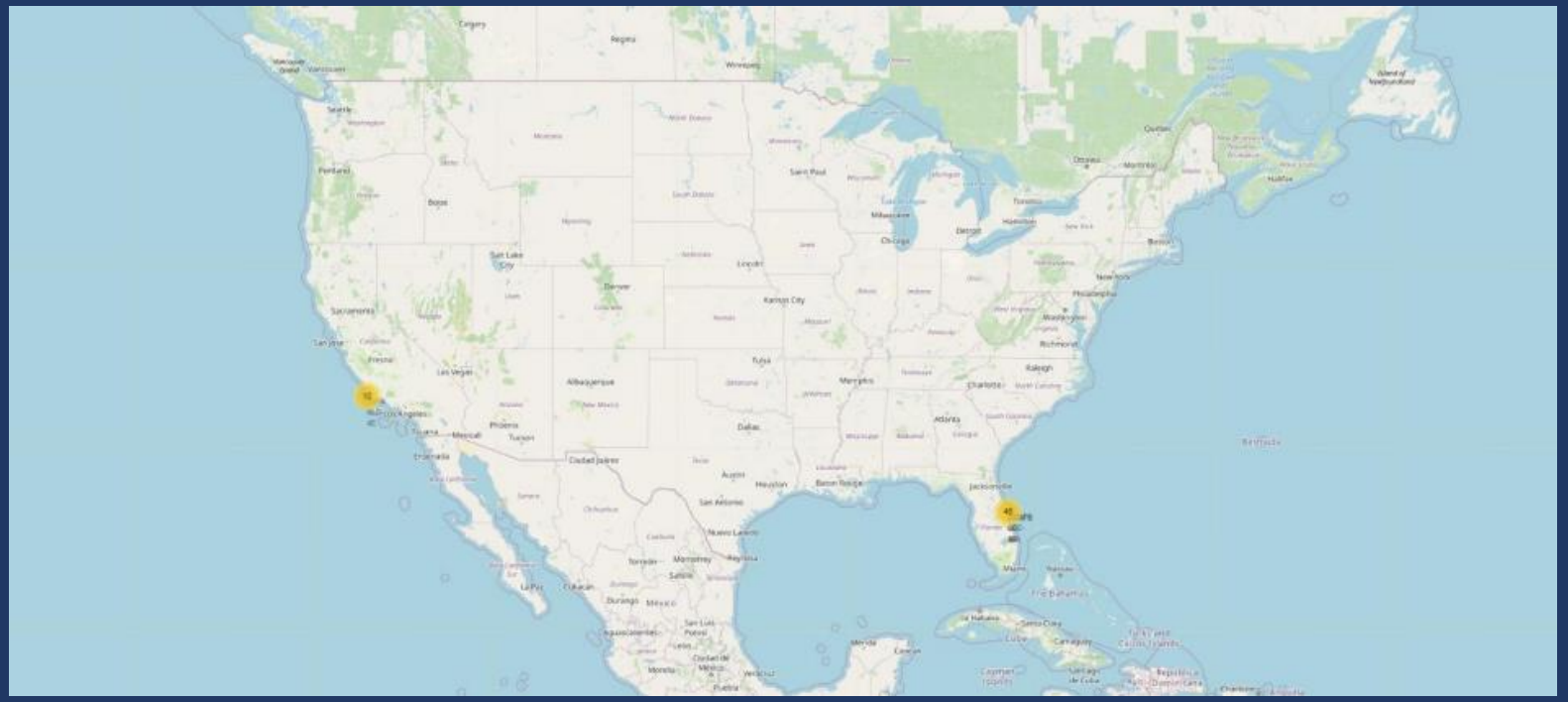
From 2010-06-04 to 2017-03-20 in descending order

```
%%sql SELECT [Landing_Outcome], count(*) as count_outcomes
FROM SPACEXTBL
WHERE DATE between '2010-06-04' and '2017-03-20' group by [Landing_Outcome] order by count_outcomes DESC;
```

Landing_Outcome	count_outcomes
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

Launch Sites

- Near the Equator: the closer the launch site is to the equator, the easier it is to launch to equatorial orbit, and more momentum you get from Earth's rotation. Rockets launched from sites near the equator get an additional natural boost due to the rotational speed of the Earth



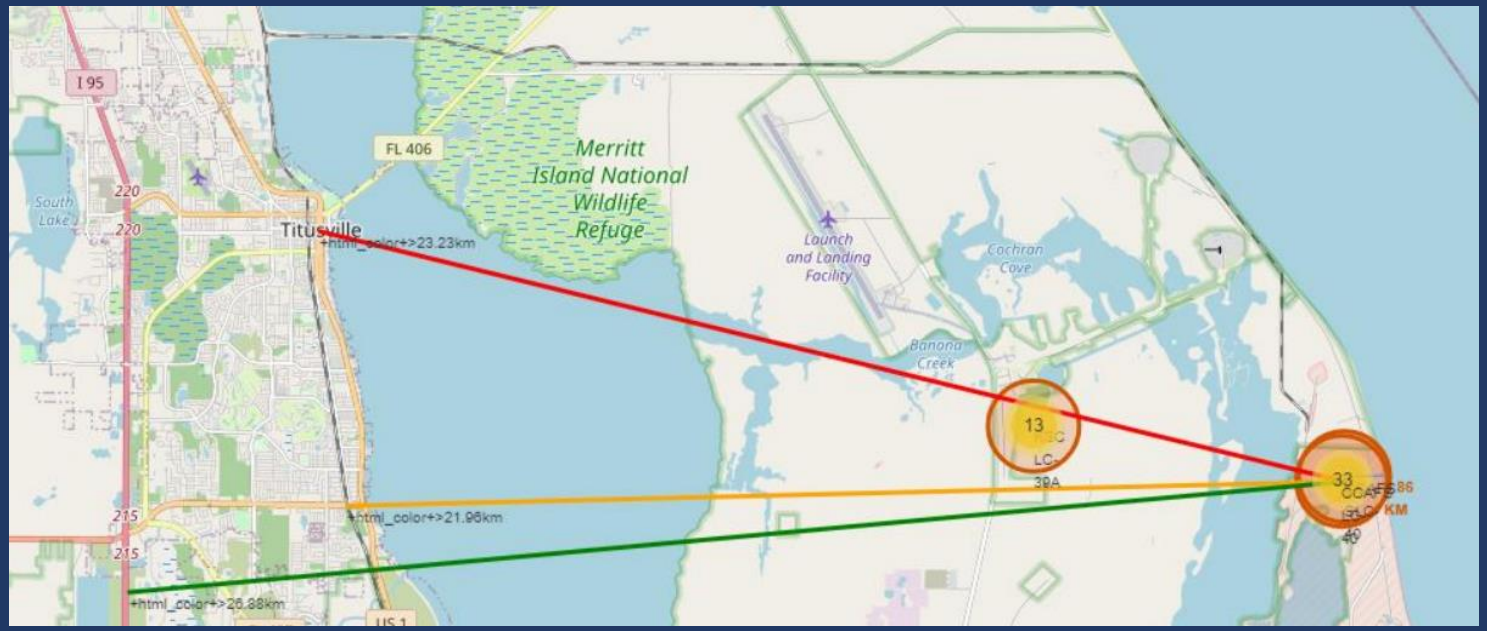
Launch Outcomes

- Green marks successful launches
- Red marks failed launches
- Launch site CCAFS SLC-40 has a 42% success rate

Distances

CCAFS SLC-40

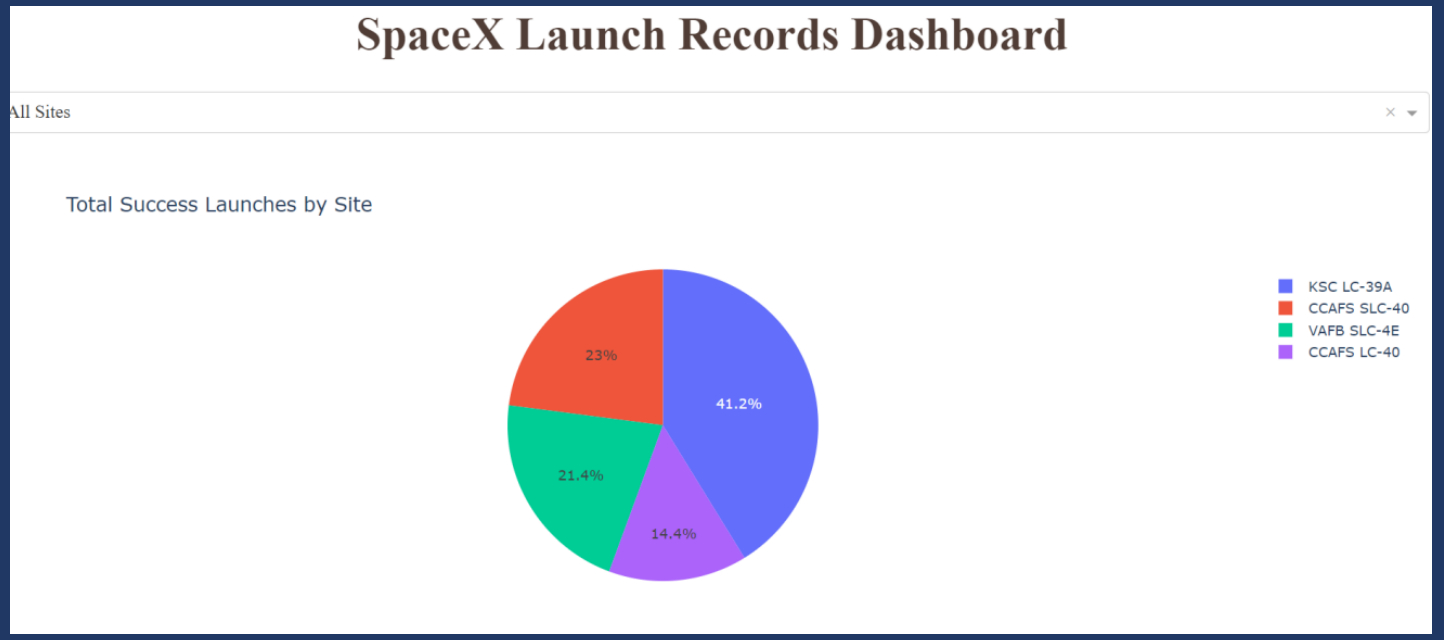
- 860m from nearest coast line
- 21,96km from nearest railway
- 23,23km from nearest city
- 26,88km from nearest highway



Launch success by Site

Success as percent of total

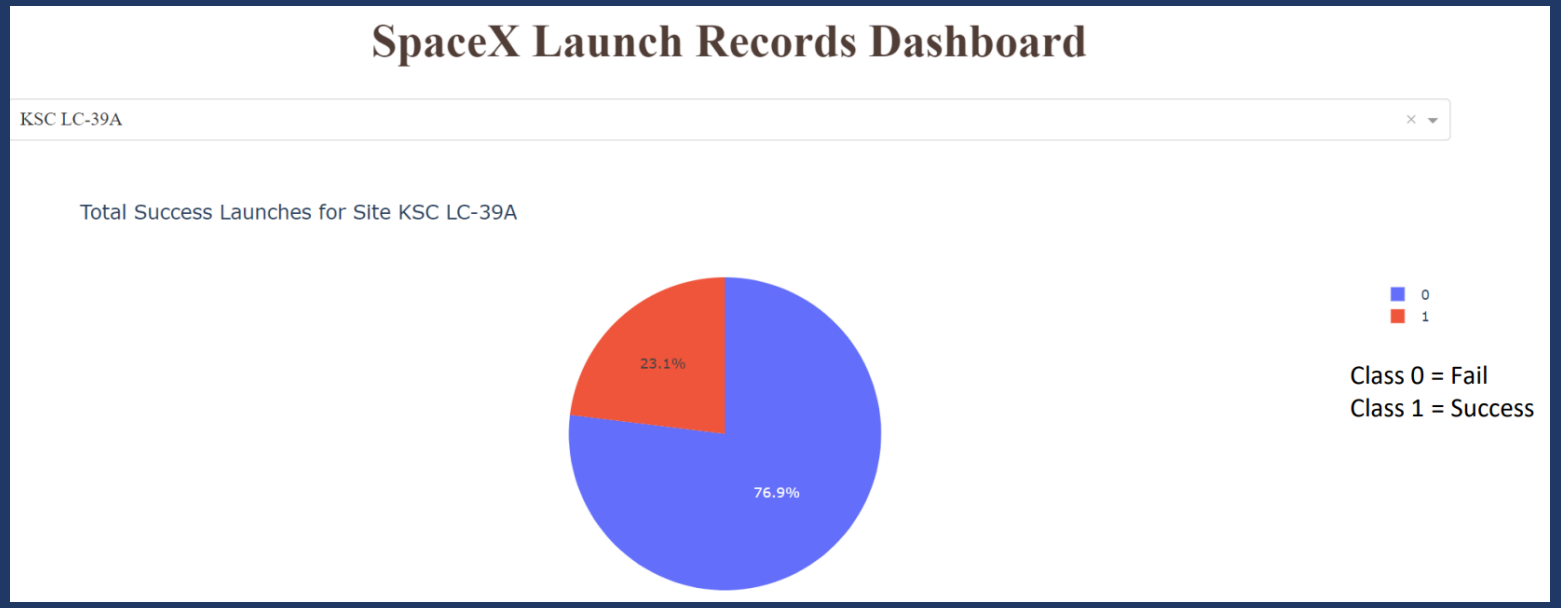
- KSC LC-39A has the most successful launches among launch site (41,2%)



Launch success (KSC LC-29A)

Success as percent of total

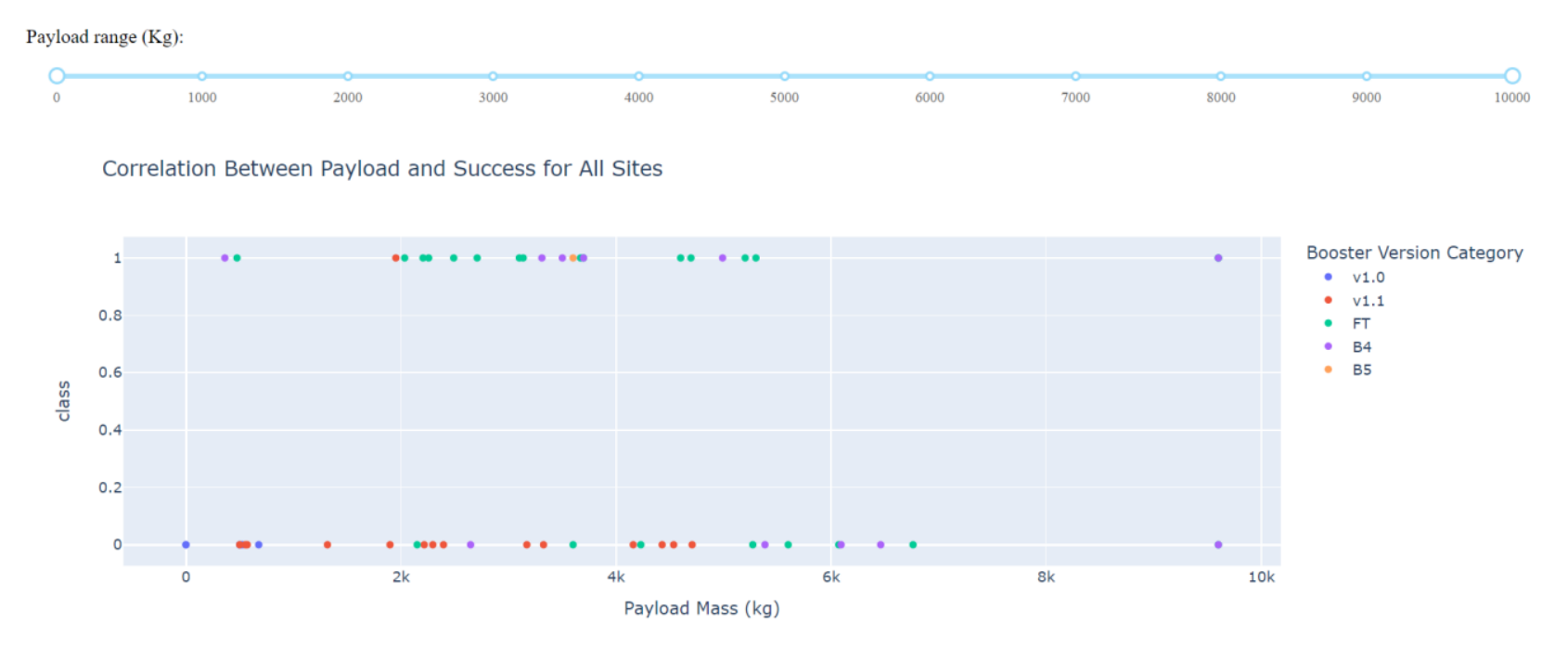
- KSC LC-29A has the highest success rate among launch sites (76,9%)



Payload Mass and Success

By Booster Version

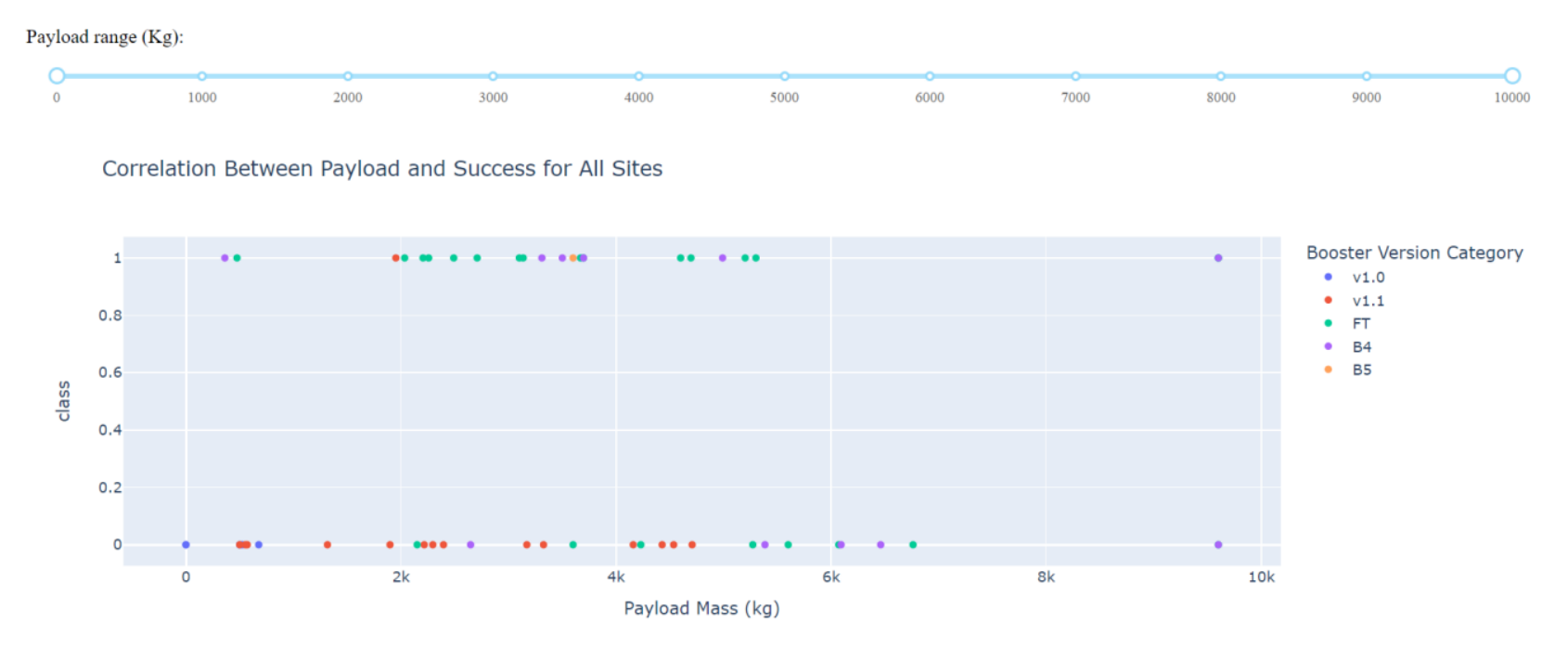
- Payloads between 2000kg and 5000kg have the highest success rate



Payload Mass and Success

By Booster Version

- Payloads between 2000kg and 5000kg have the highest success rate



Predictive Analysis - Classification

Accuracy

- All the models performed at about the same level and had the same scores and accuracy. This is likely due to the small dataset. The Decision Tree model slightly outperformed the rest when looking at the average of all cv folds for a single combination of the parameters

	LogReg	SVM	Tree	KNN
Jaccard_Score	0.800000	0.800000	0.600000	0.800000
F1_Score	0.888889	0.888889	0.750000	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

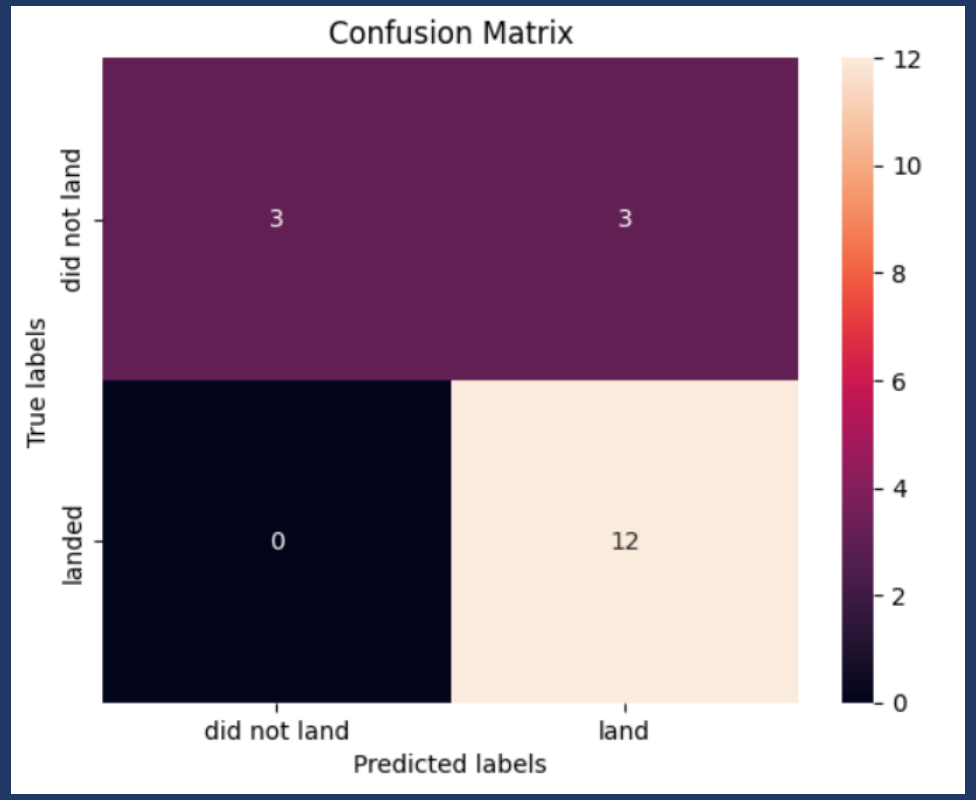
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

bestalgorithm = max(models, key=models.get)
print('Best model is', bestalgorithm,'with a score of', models[bestalgorithm])
if bestalgorithm == 'DecisionTree':
    print('Best params is :', tree_cv.best_params_)
if bestalgorithm == 'KNeighbors':
    print('Best params is :', knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best params is :', logreg_cv.best_params_)
if bestalgorithm == 'SupportVector':
    print('Best params is :', svm_cv.best_params_)
```

```
Best model is DecisionTree with a score of 0.9035714285714287
Best params is : {'criterion': 'entropy', 'max_depth': 6, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_spli
t': 5, 'splitter': 'random'}
```

Predictive Analysis – Confusion Matrices

Performance Summary

- A confusion matrix summarizes the performance of a classification algorithm
- All the confusion matrices were identical



Conclusion

- Model Performance: The models performed similarly on the test set with the Decision Tree Model slightly outperforming

- As the number of flights increases, the rate of success at a launch site increases, with most early flights being unsuccessful (with more experience, more success)

- Orbits types ES-L1, GEO, HEO and SSO have 100% success rate
  GEO, HEO and ES-L1 had only 1 flight
  SSO had 5 successful flights

- The launch site KSC LC-39A had the most successful launches, with 41,7% of the total successful launches, and also the highest rate of succesful launches, with a 76,9% success rate

- The success for massive payloads (over 4000kg) is lower than for lower payloads

IBM DATA SCIENCE CAPSTONE PROJECT

LPNM