IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Yi-Hsuan Lin
2023.09.30

Outline

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

Executive Summary

- The research attempts to identify the factors for a successful rocket landing. To make this determination, the following methodologies were used:
- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data to create a success/fail outcome variable
- Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number, and yearly trend
- Analyze the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total number of successful and failed outcomes
- Explore launch site success rates and proximity to geographical markers
- Visualize the launch sites with the most successful payload ranges
- Build Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN)

Introduction

SpaceX strives to make space travel affordable for everyone. Its accomplishments include sending spacecraft to the International Space Station, launching a satellite constellation that provides internet access, and sending manned missions to space. SpaceX can do this because the rocket launches are relatively inexpensive ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket. Other providers, which are not able to reuse the first stage, cost upwards of $165 million each. By determining if the first stage will land, we can determine the price of the launch. To do this, we can use public data and machine learning models to predict whether SpaceX – or a competing company – can reuse the first stage.

Section 1

Methodology

Methodology

- Collect data using SpaceX REST API and web scraping techniques
- Wrangle data to create a success/fail outcome variable
- Explore data with data visualization techniques, considering the following factors: payload, launch site, flight number, and yearly trend
- Analyze the data with SQL, calculating the following statistics: total payload, payload range for successful launches, and total # of successful and failed outcomes
- Explore launch site success rates and proximity to geographical markers
- Visualize the launch sites with the most successful payload ranges
- Build Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN)

Data Collection – SpaceX API

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

Data Collection - Scraping

- Request data (Falcon 9 launch data) from Wikipedia
- Create BeautifulSoup object from HTML response
- Extract column names from the HTML table header
- Collect data from parsing HTML tables
- Create a dictionary from the data
- Create a data frame from the dictionary
- Export data to CSV file

Data Wrangling

Convert outcomes into 1 for a successful landing and 0 for an unsuccessful landing

- Perform EDA and determine data label
- Calculate the number of launches for each site, occurrence of orbit, and occurrence of mission outcome per orbit type.
- Create a binary landing outcome column
- Export data to CSV file

EDA with Data Visualization

Charts:

- Flight number vs. Payload
- Flight number vs. Launch Site
- Payload Mass vs. Launch Site
- Payload Mass vs. Orbit type

Analysis

- View the 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 the launch site begins with ‘CCA’
- Total payload mass carried by boosters launched by NASA (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of first successful landing on a ground pad
- Names of boosters that had success landing on drone ships and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions • Names of booster versions that have carried the max payload
- Failed landing outcomes on drone ship, their booster version, and launch site for the months in the year 2015
- Count of landing outcomes between 2010-06-04 and 2017-03-20 (desc)

Build an Interactive Map with Folium

Markers Indicating Launch Sites

- Added blue circle at NASA Johnson Space Center's coordinate with a popup label showing its name using its latitude and longitude coordinates
- Added red circles at all launch site coordinates with a popup label showing its name using its name using its latitude and longitude coordinates Map with Folium

Colored Markers of Launch Outcomes

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

Distances Between a Launch Site to Proximities

- Added colored lines to show distance between launch site CCAFS SLC40 and its proximity to the nearest coastline, railway, highway, and city

Build a Dashboard with Plotly Dash

Dropdown List with Launch Sites

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

Dashboard with Plotly Dash Slider of Payload Mass Range

- Allow user to select payload mass range Pie Chart

Showing Successful Launches

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

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

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

Predictive Analysis (Classification)

Charts

- Create NumPy array from the Class column
- Standardize the data with StandardScaler.
- Fit and transform the data.
- Split the data using train_test_split
- Create a GridSearchCV object with cv=10 for parameter optimization
- Apply GridSearchCV on different algorithms: logistic regression (LogisticRegression()), support vector machine (SVC()), decision tree (DecisionTreeClassifier()), K-Nearest Neighbor (KNeighborsClassifier())
- 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

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 away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

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

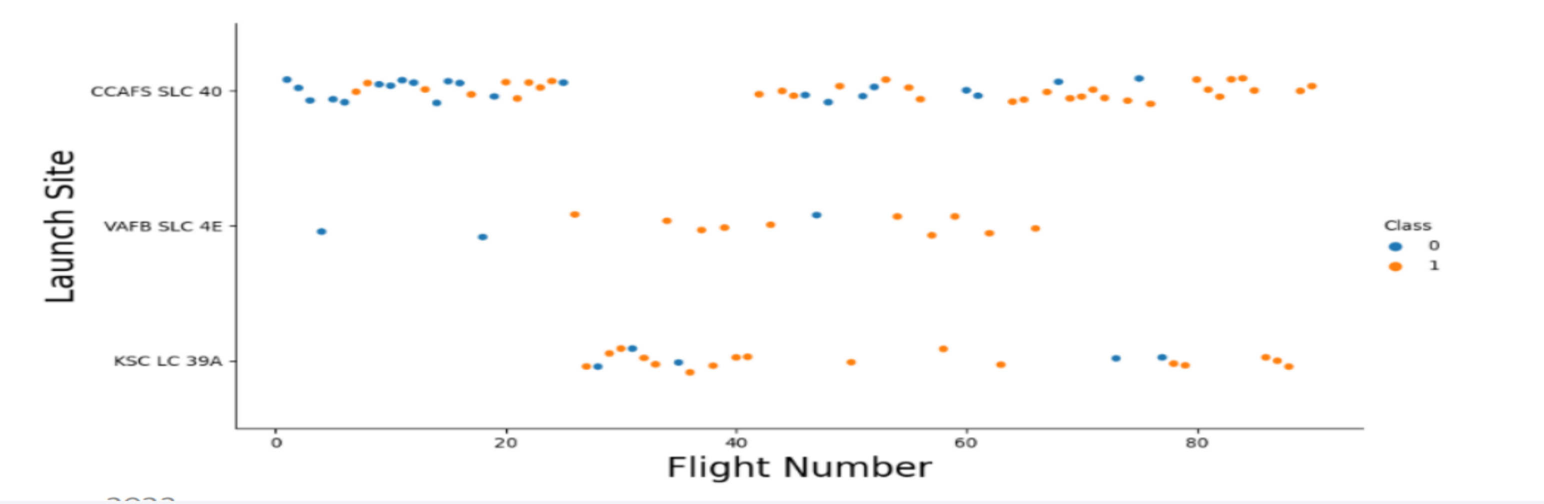
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

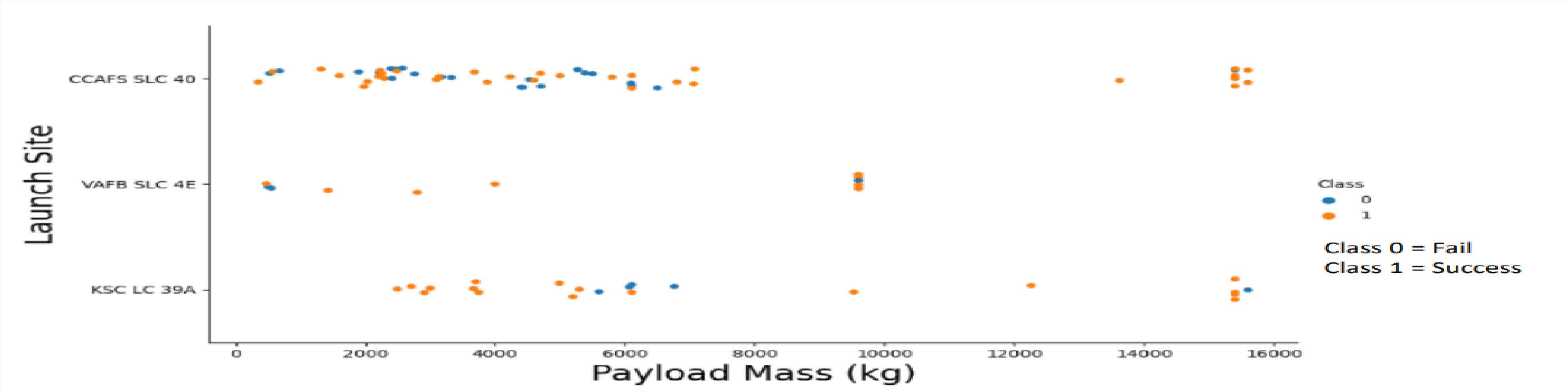
- Earlier flights had a lower success rate (blue = fail)
- Later flights had a higher success rate (orange = success)
- Around half of the launches were from the CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher success rate



Payload vs. Launch Site

Exploratory Data Analysis

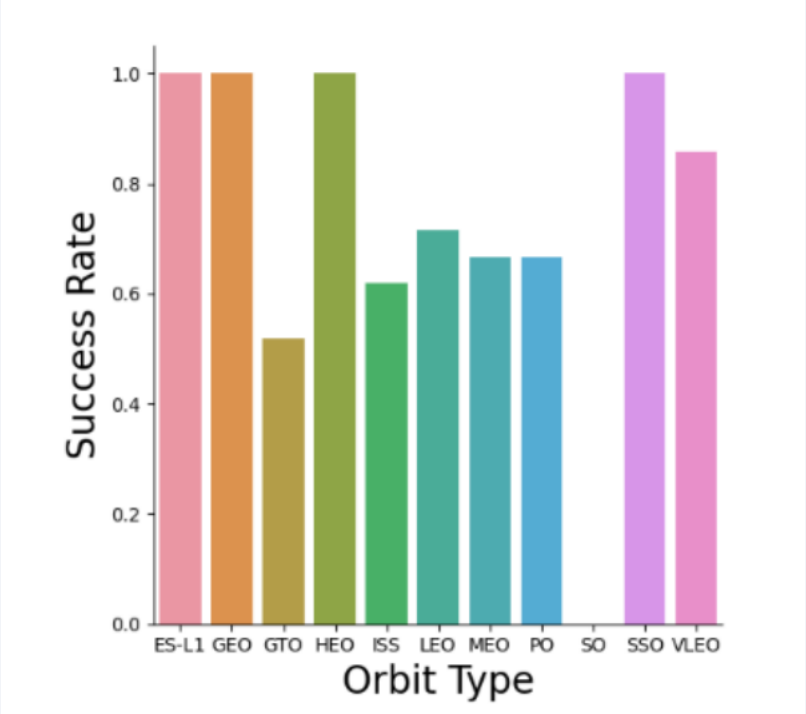
- Typically, the higher the payload mass (kg), the higher the success rate
- Most launches with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



Success Rate vs. Orbit Type

Exploratory Data Analysis

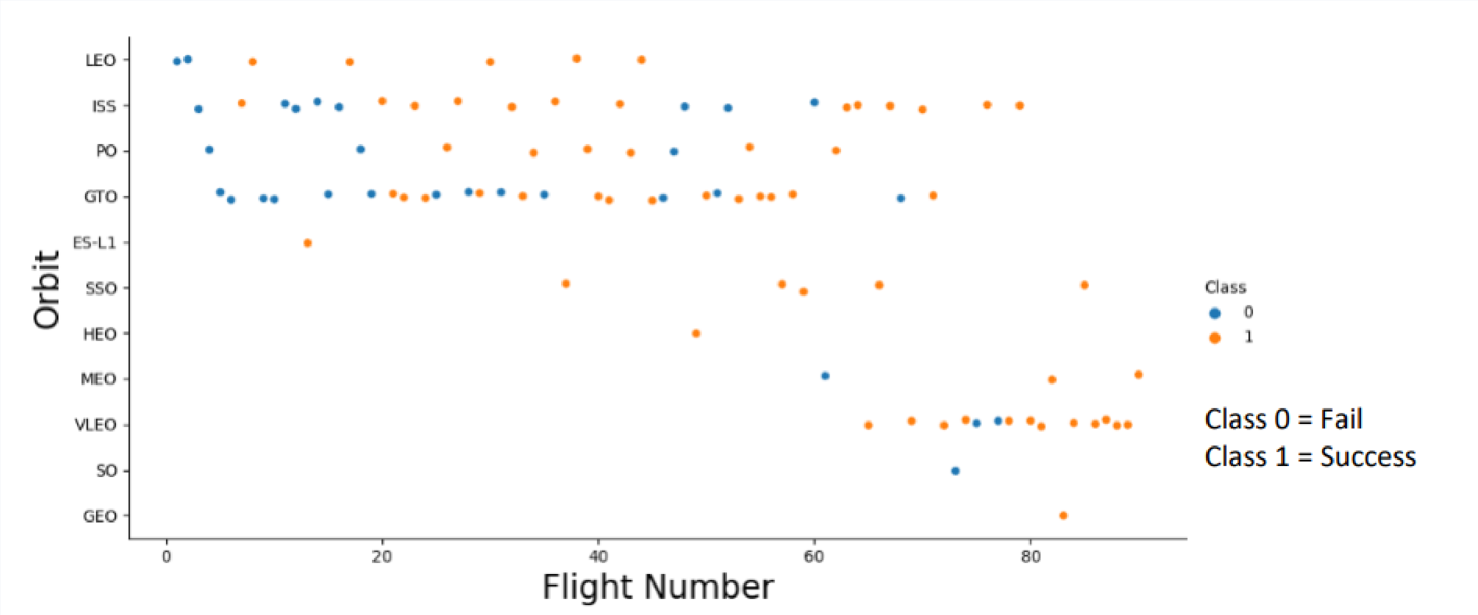
- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



Flight Number vs. Orbit Type

Exploratory Data Analysis

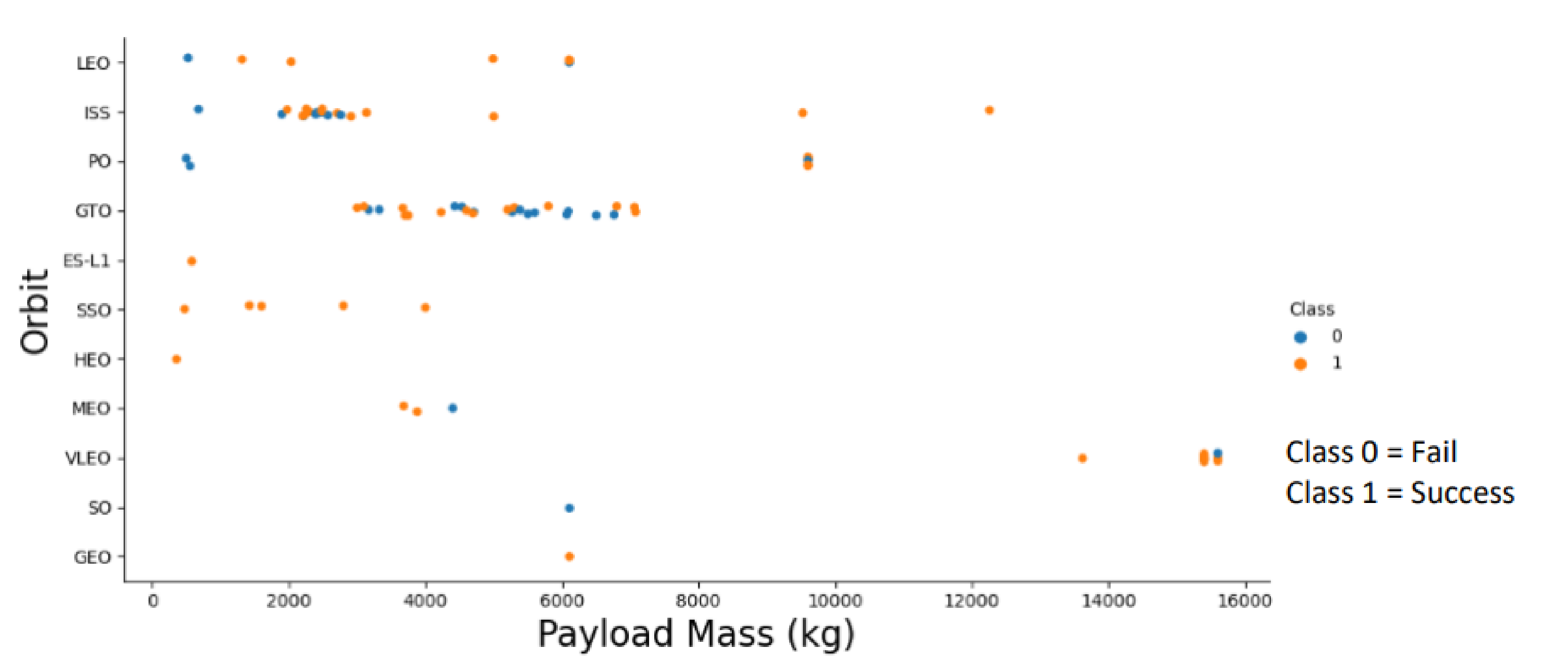
- The success rate typically increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit, however, does not follow this trend



Payload vs. Orbit Type

Exploratory Data Analysis

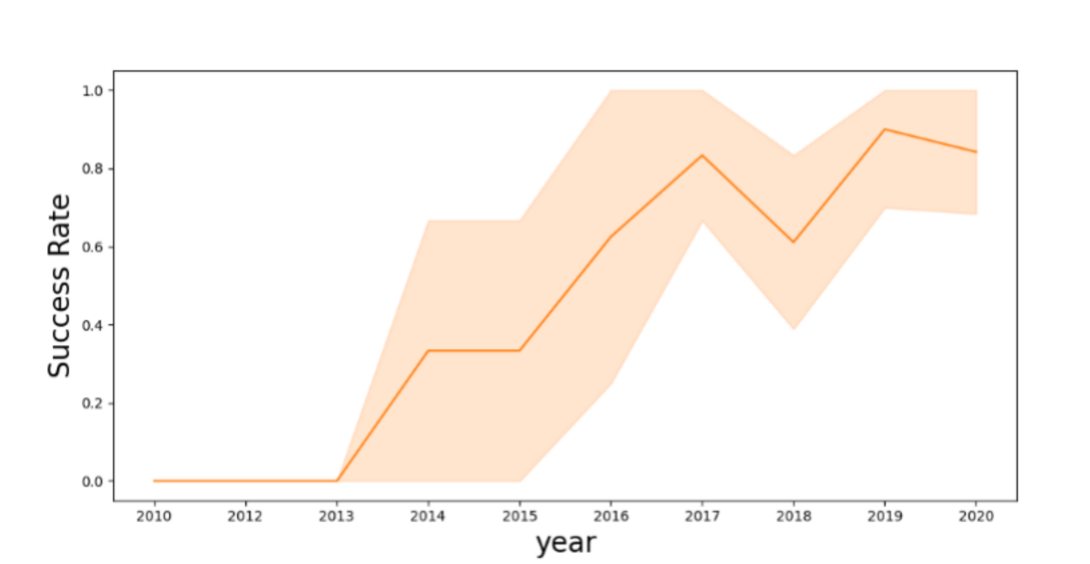
- Heavy payloads are better with LEO, ISS, and PO orbits
- The GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

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



All Launch Site Names

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

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-ab1a4348f4a4.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:32286/BLUDB
   sqlite:///my_data1.db
Done.
```

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

Total Payload Mass

- Total Payload Mass
- 45,596 kg (total) carried by boosters launched by NASA (CRS)
- Average Payload Mass
- 45,596 kg (average) carried by boosters launched by NASA (CRS)

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4l
   sqlite:///my_data1.db
Done.
```

1
45596

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4
   sqlite:///my_data1.db
Done.
```

1
2928

First Successful Ground Landing Date

2015-12-22

```
%sql SELECT MIN(DATE) \
FROM SPACEXTBL \
WHERE LANDING__OUTCOME = 'Success (ground pad)'

 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b
   sqlite:///my_data1.db
Done.
```

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster mass greater than 4,000 but less than 6,000
- JSCAT-14, JSCAT-16, SES-10, SES-11 / EchoStar 105 B

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

```
 * ibm_db_sa://yyy33800:***@1bbf73c5-d84a-4bb0-85b9-
   sqlite:///my_data1.db
Done.
```

payload
JCSAT-14
JCSAT-16
SES-10
SES-11 / EchoStar 105

Total Number of Successful and Failure Mission Outcomes

- 1 Failure in Flight
- 99 Success
- 1 Success (payload status unclear)

```
%sql SELECT MISSION_OUTCOME, COUNT(*) as total_number \
FROM SPACEXTBL \
GROUP BY MISSION_OUTCOME;
```

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

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

Boosters Carried Maximum Payload

- Carrying Max Payload
- 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

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

```
 * 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

2015 Launch Records

- Showing month, date, booster version, launch site and landing outcome

```
%sql SELECT substr(Date,4,2) as month, DATE,BOOSTER_VERSION, LAUNCH_SITE, [Landing _Outcome] \
FROM SPACEXTBL \
where [Landing _Outcome] = 'Failure (drone ship)' and substr(Date,7,4)='2015';
```

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

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

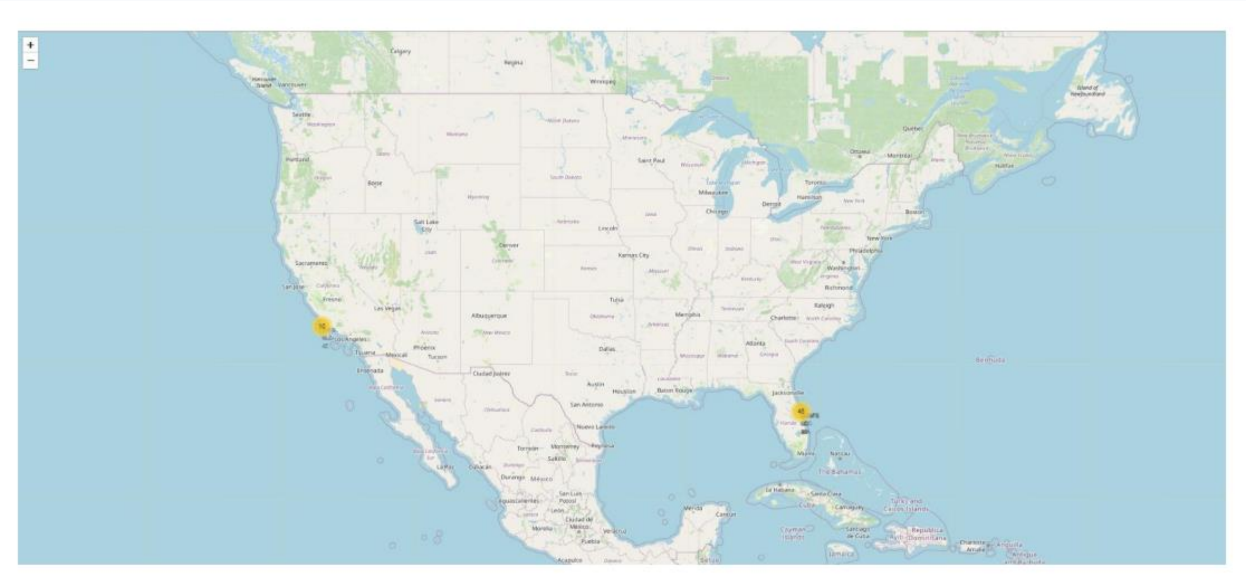
Section 3

Launch Sites Proximities Analysis

Launch site

With Markers

- Near Equator: the closer the launch site to the equator, the easier it is to launch to equatorial orbit, and the more help you get from Earth's rotation for a prograde orbit. Rockets launched from sites near the equator get an additional natural boost - due to the rotational speed of earth - that helps save the cost of putting in extra fuel and boosters.



Launch Outcomes

At Each Launch Site

Outcomes:

Green markers for successful launches

Red markers for unsuccessful launches

Launch site CCAFS SLC-40 has a 3/7 success rate (42.9%)

Distance to Proximities

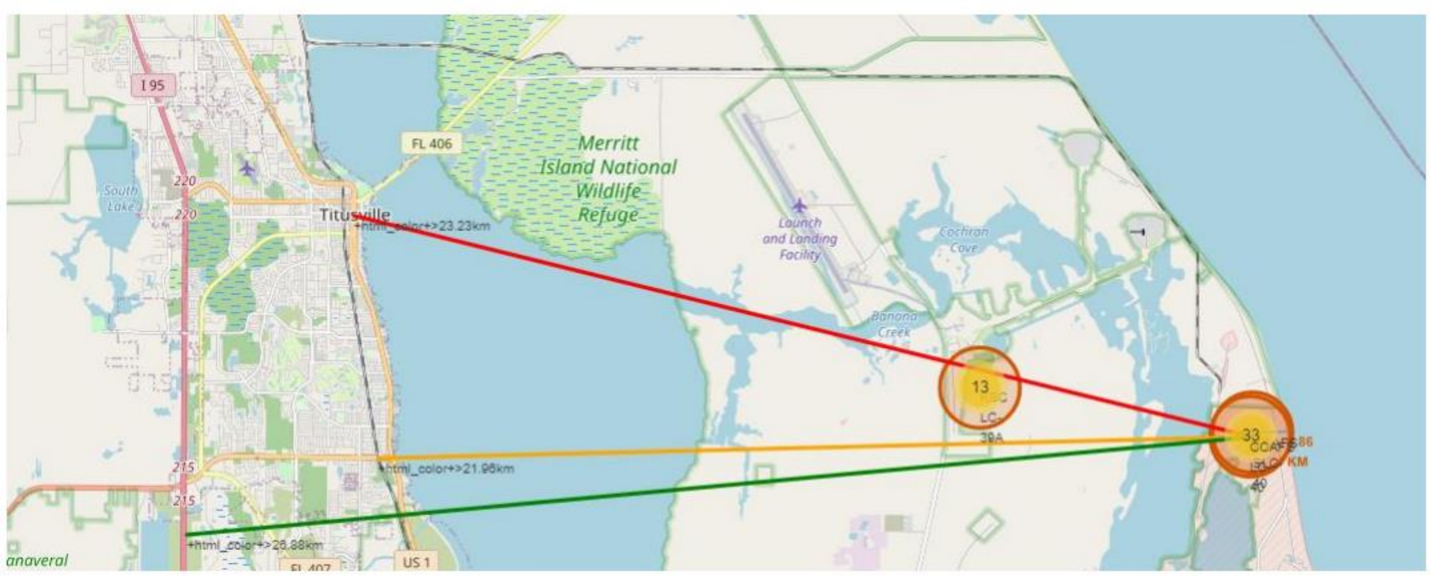
CCAFS SLC-40

.86 km from nearest coastline

21.96 km from nearest railway

23.23 km from nearest city

26.88 km from nearest highway



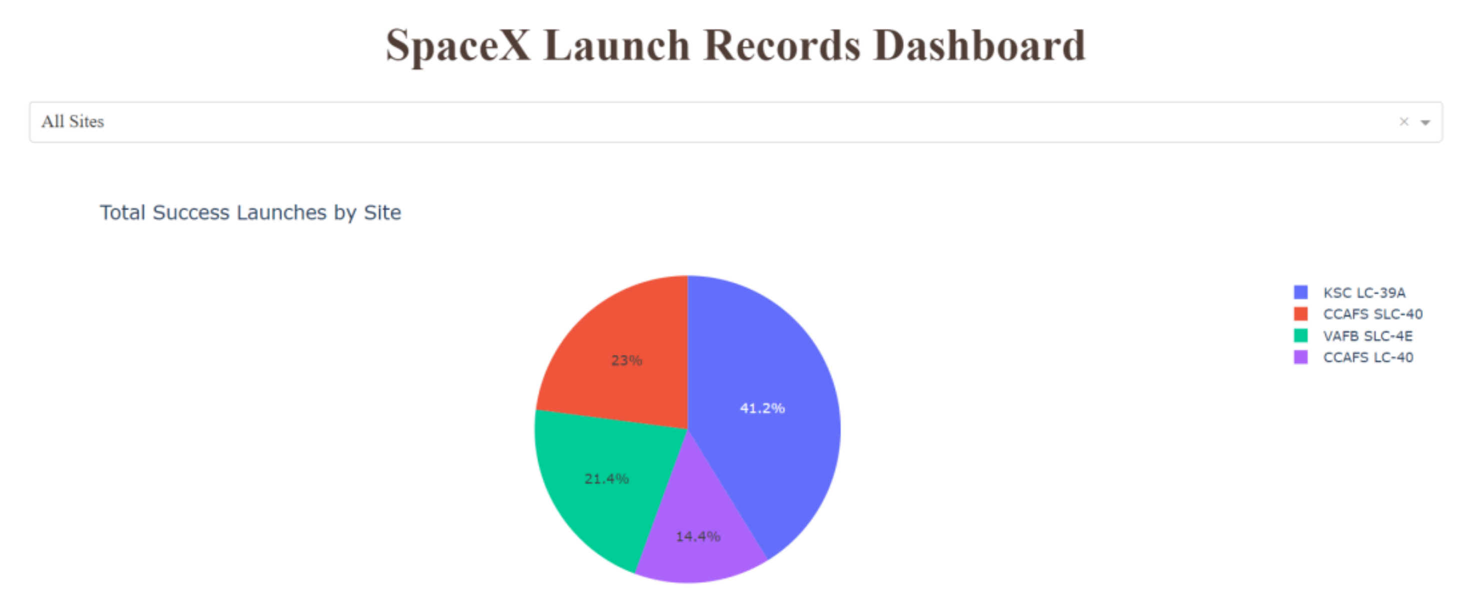
Section 4

Build a Dashboard with Plotly Dash

Launch successful by site

Success as Percent of Total

KSC LC-39A has the most successful launches amongst launch sites (41.2%)

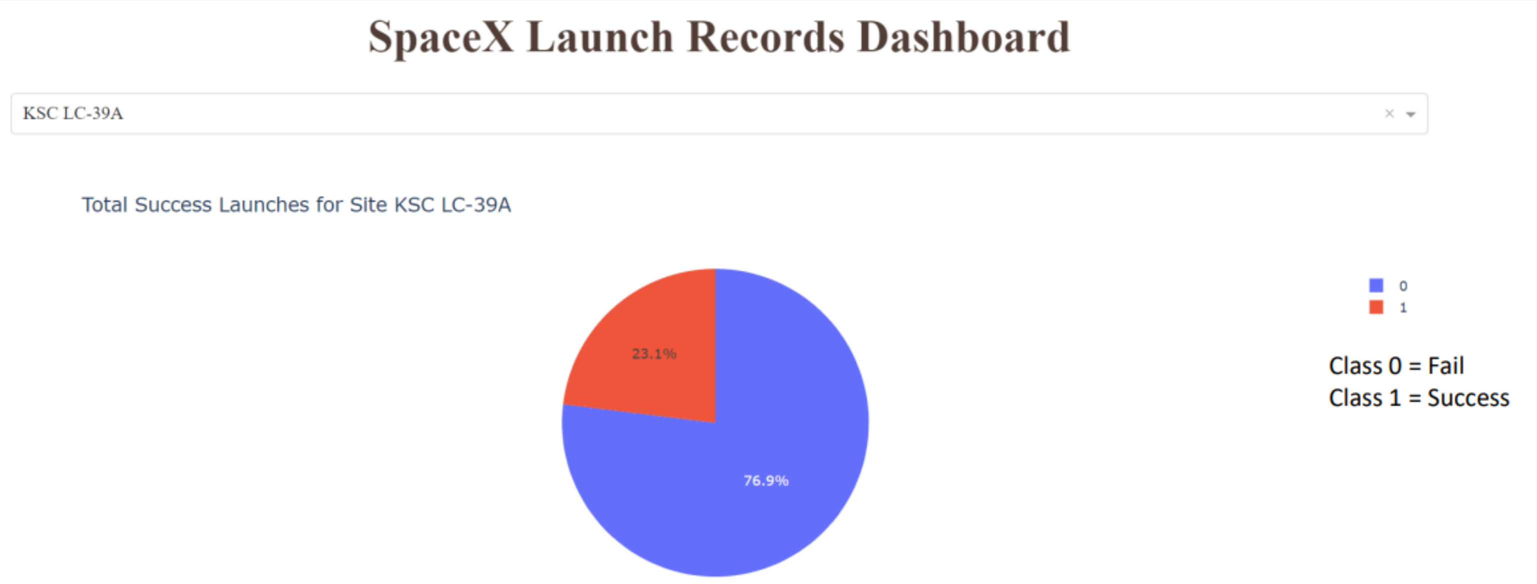


Launch Success (KSC LC-29A)

Success as Percent of Total

KSC LC-39A has the highest success rate amongst launch sites (76.9%)

10 successful launches and 3 failed launches

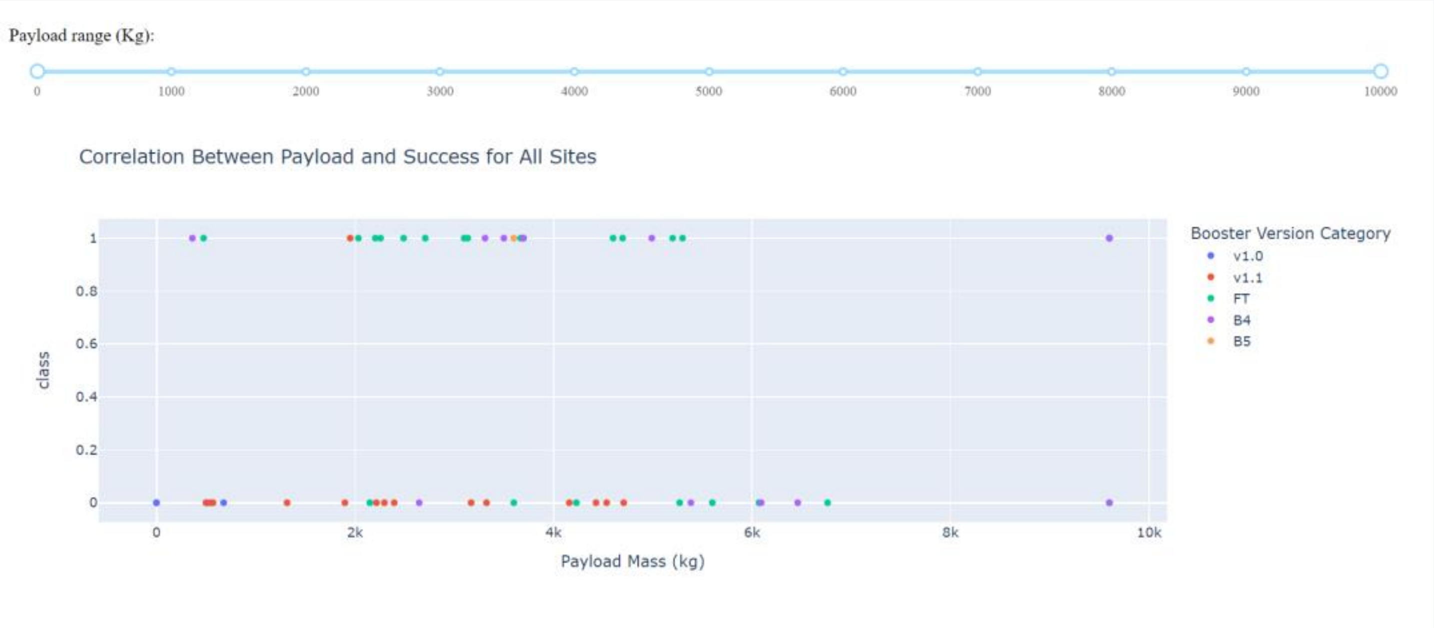


Payload Mass and Success

By Booster Version

Payloads between 2,000 kg and 5,000 kg have the highest success rate

1 indicating successful outcome and 0 indicating an unsuccessful outcome



Section 5

Predictive Analysis (Classification)

Classification Accuracy

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 .best_score_

.best_score_ is the average of all cv folds for a single combination of the parameters

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

```python
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 0.9017857142857142
Best params is : {'criterion': 'gini', 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 4, 'min_samples_split': 10, 'splitter': 'random'}
```

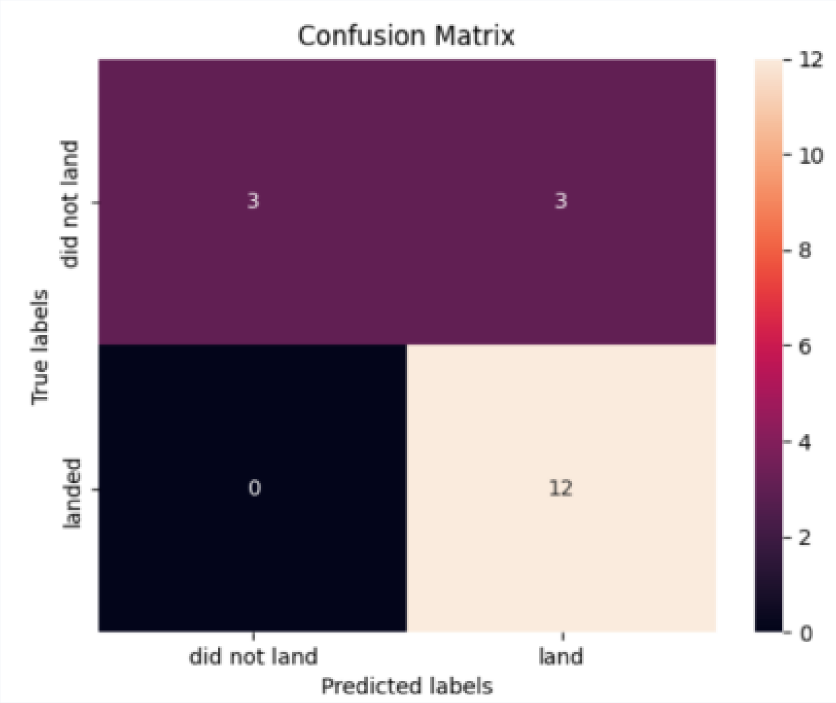
Confusion Matrix

Performance Summary

A confusion matrix summarizes the performance of a classification algorithm

All the confusion matrices were identical

The fact that there are false positives (Type 1 error) is not good



Conclusions

Research

- Model Performance: The models performed similarly on the test set with the decision tree model slightly outperforming
- Equator: Most of the launch sites are near the equator for an additional natural boost - due to the rotational speed of earth - which helps save the cost of putting in extra fuel and boosters
- Coast: All the launch sites are close to the coast
- Launch Success: Increases over time
- KSC LC-39A: Has the highest success rate among launch sites. Has a 100% success rate for launches less than 5,500 kg
- Orbits: ES-L1, GEO, HEO, and SSO have a 100% success rate
- Payload Mass: Across all launch sites, the higher the payload mass (kg), the higher the success rate

Things to Consider

- Dataset: A larger dataset will help build on the predictive analytics results to help understand if the findings can be generalizable to a larger data set
- Feature Analysis / PCA: Additional feature analysis or principal component analysis should be conducted to see if it can help improve accuracy
- XGBoost: Is a powerful model which was not utilized in this study. It would be interesting to see if it outperforms the other classification models

Thank you!