Winning Space Race with Data Science

<Name>
<Date>

Table Of Contents

Executive Summary

This research aims to identify the factors for successful rocket landings. The following methodologies were employed:

- Data collection using SpaceX REST API and web scraping techniques
- Data wrangling to create a success/fail outcome variable
- Data exploration using data visualization techniques, considering factors such as payload, launch site, flight number, and yearly trends
- Data analysis using SQL to calculate statistics including total payload, payload range for successful launches, and the total number of successful and failed outcomes
- Exploration of launch site success rates and their proximity to geographical markers
- Visualization of launch sites with the highest success rates and successful payload ranges
- Building models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree, and K-nearest neighbor (KNN) algorithms.

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

Visualization/Analytics:

- Most launch sites are near the equator, and all are close to the coast

Predictive Analytics:

- All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

Background

- SpaceX, a prominent player in the space industry, is committed to making space travel accessible to a wider audience. Their achievements include delivering spacecraft to the international space station, deploying a satellite network for global internet connectivity, and successfully conducting manned missions to space. One of the key factors enabling SpaceX to achieve these milestones is the cost-efficiency of their rocket launches. By innovatively reusing the first stage of their Falcon 9 rocket, SpaceX is able to significantly reduce launch expenses to approximately $62 million per launch. In contrast, other providers who do not employ stage reuse face costs exceeding $165 million per launch.
- The successful landing of the first stage plays a crucial role in determining the overall launch price. To ascertain whether the first stage can be reused, public data and machine learning models can be utilized to predict the outcome. This analysis enables SpaceX, as well as their competitors, to evaluate the feasibility of first stage reuse and make informed decisions regarding launch costs and strategies.

Explore

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing (binary classification)

Section 1

Methodology

Methodology

Here are the steps involved in the project:

1. Collect data: Gather data from SpaceX using their REST API and employ web scraping techniques to obtain additional relevant information.

2. Data wrangling: Filter the collected data, handle missing values, and apply one-hot encoding or any other necessary data transformations to prepare the data for analysis and modeling.

3. Explore data: Perform exploratory data analysis (EDA) using SQL queries and data visualization techniques. Analyze various factors and relationships such as launch outcomes, payload, launch site, and other relevant variables.

4. Visualize data: Utilize libraries like Folium and Plotly Dash to create interactive and visually appealing visualizations of the data. This helps in better understanding and communication of the findings.

5. Build prediction models: Employ classification models to predict landing outcomes. Use machine learning techniques to train and evaluate these models. Consider models such as logistic regression, support vector machines (SVM), decision trees, and other suitable algorithms.

6. Model tuning and evaluation: Fine-tune the models by adjusting parameters and hyperparameters. Evaluate the performance of each model using appropriate metrics. Compare the models to identify the best-performing one based on accuracy, precision, recall, or other relevant criteria.

By following these steps, the project aims to collect, preprocess, explore, visualize, and model the data to predict landing outcomes accurately.

Data Collection – SpaceX API

Steps

1. **Request data** from SpaceX API (rocket launch data)
2. **Decode response** using .json() and convert to a dataframe using .json_normalize()
3. **Request information** about the launches from SpaceX API using custom functions
4. **Create dictionary** from the data
5. **Create dataframe** from the dictionary
6. **Filter dataframe** to contain only Falcon 9 launches
7. **Replace missing values** of Payload Mass with calculated .mean()
8. **Export data** to csv file

Data Collection - Scraping

Steps

1. **Request data** (Falcon 9 launch data) from Wikipedia
2. **Create BeautifulSoup** object from HTML response
3. **Extract column names** from HTML table header
4. **Collect data** from parsing HTML tables
5. **Create dictionary** from the data
6. **Create dataframe** from the dictionary
7. **Export data** to csv file

Data Wrangling

1. Perform EDA and determine data labels: Explore the data to understand its structure, variables, and any missing or inconsistent values. Identify the relevant labels or categories for analysis.
2. Calculate key metrics: Calculate the following metrics:
   1. Number of launches for each launch site: Count the occurrences of launches for each specific launch site.
   2. Number and occurrence of orbit types: Count the number of occurrences for each orbit type.
   3. Number and occurrence of mission outcomes per orbit type: Count the occurrences of mission outcomes (success or failure) for each orbit type.
3. Create a binary landing outcome column: Create a new column called "Landing Outcome" to represent the landing outcome as a binary variable. Assign the following values:
   1. False Ocean: Represents an unsuccessful landing in a specific region of the ocean.
   2. True RTLS: Indicates a successful landing on a ground pad.
   3. False RTLS: Represents an unsuccessful landing on a ground pad.
   4. True ASDS: Indicates a successful landing on a drone ship.
   5. False ASDS: Represents an unsuccessful landing on a drone ship.
4. Convert outcomes into binary values: Convert the landing outcomes into binary values, where 1 represents a successful landing and 0 represents an unsuccessful landing.
5. Export data to CSV file: Save the processed and wrangled data into a CSV file for further analysis and modeling.

Landing Outcome:

- Note that the landing was not always successful.
- True Ocean: Indicates a mission outcome with a successful landing in a specific region of the ocean.

By performing these data wrangling steps, you will have processed the data, calculated relevant metrics, created a binary landing outcome column, and converted outcomes into binary values for further analysis and modeling.

EDA with Data Visualization

Charts

- Flight Number vs. Payload
- Flight Number vs. Launch Site
- Payload Mass (kg) vs. Launch Site
- Payload Mass (kg) vs. Orbit type

Analysis

- Interpret the scatter plots to identify any correlations or patterns between variables, such as flight numbers and payload mass.
- Analyze the bar charts to compare and understand the distribution and relationships among discrete categories, such as flight numbers across launch sites or payload mass across orbit types.
- These visualizations can provide valuable insights and serve as a basis for further analysis and machine learning modeling.

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 (CRS)
- Average payload mass carried by booster version F9 v1.1.

List:

- Date of first successful landing on ground pad
- Names of boosters which had success landing on drone ship and have payload mass greater than 4,000 but less than 6,000
- Total number of successful and failed missions
- Names of booster versions which 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 sites coordinates with a popup label showing its name using its name using its latitude and longitude coordinates

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

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 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

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

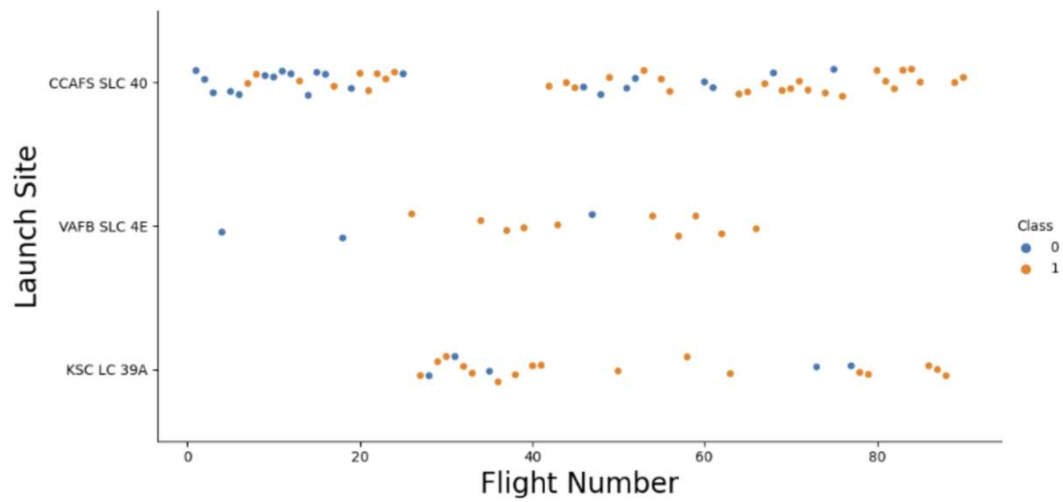
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

Exploratory Data Analysis

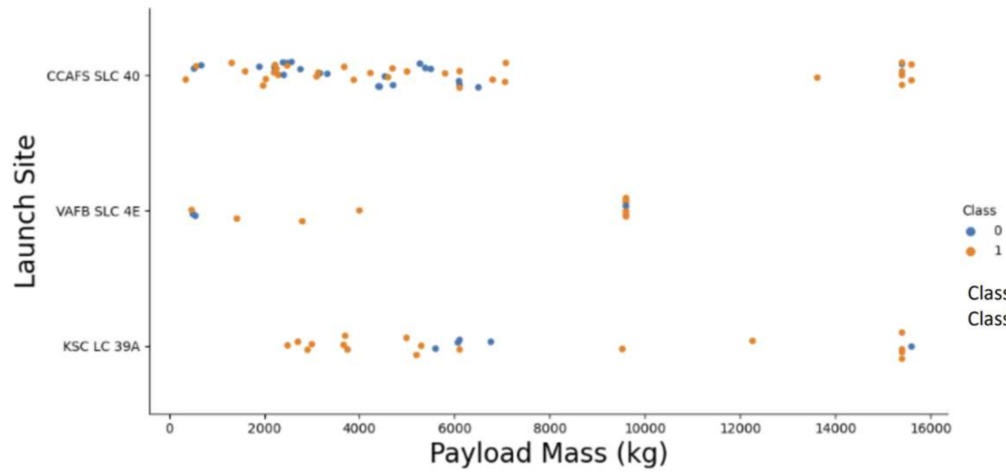
- **Earlier flights** had a **lower success rate** (**blue = fail**)
- **Later flights** had a **higher success rate** (**orange = success**)
- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- We can infer that new launches have a higher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



Payload vs. Launch Site

Exploratory Data Analysis

- Typically, the **higher** the **payload mass** (kg), the **higher** the **success rate**
- Most launces 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

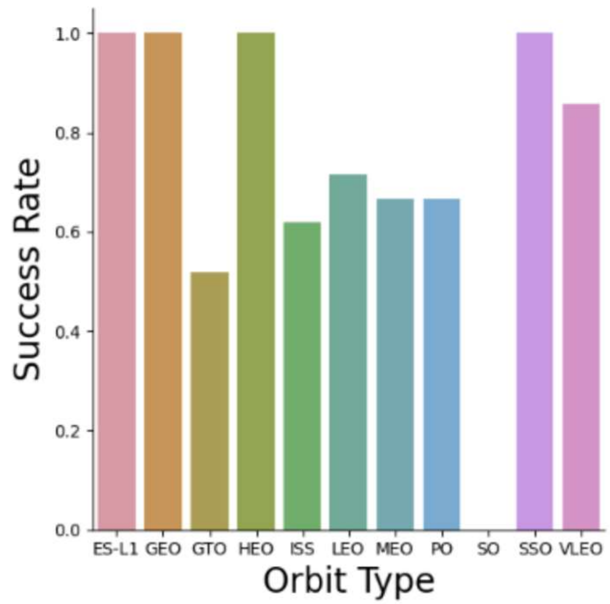


Class 0 = Fail
Class 1 = Success

Success Rate vs. Orbit Type

Exploratory Data Analysis

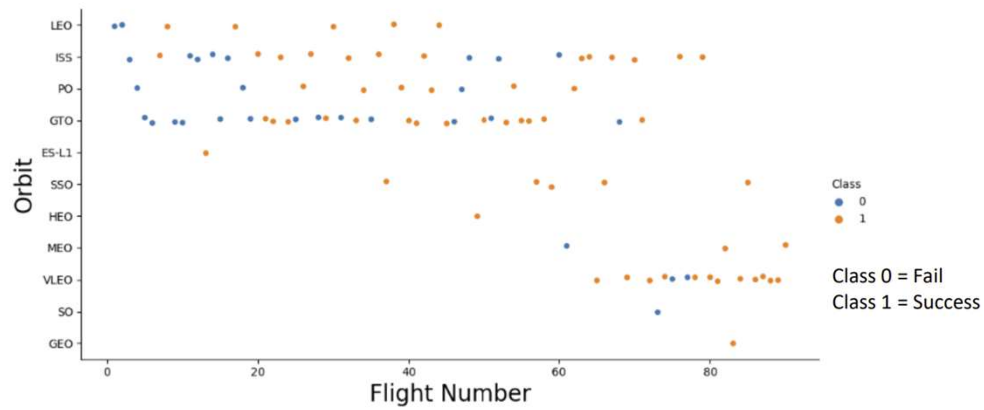
- **100% Success Rate**: ES-L1, GEO, HEO and SSO
- **50%-80% Success Rate**: GTO, ISS, LEO, MEO, PO
- **0% Success Rate**: SO



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

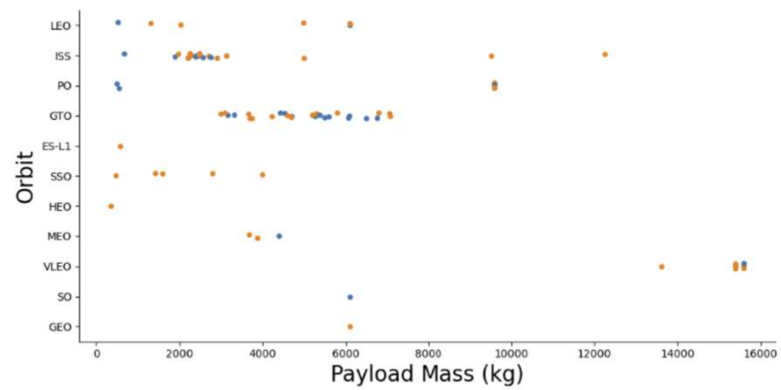


Payload vs. Orbit Type

Payload vs. Orbit

Exploratory Data Analysis

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

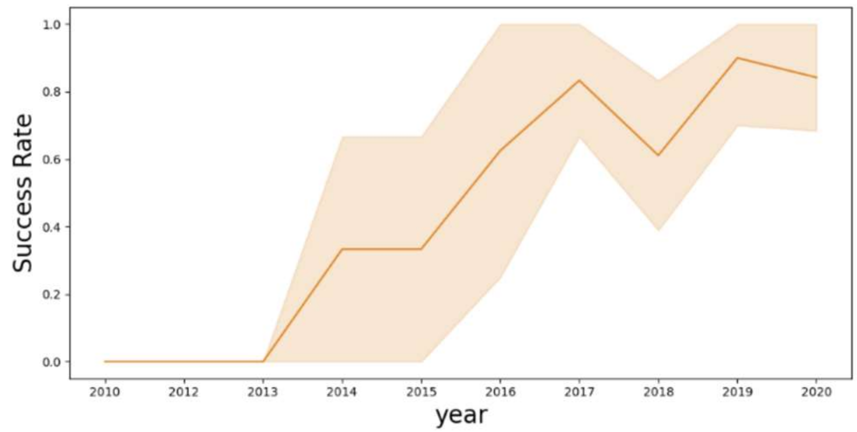


Class 0 = Fail
Class 1 = Success

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



Launch Site Names Begin with 'CCA'

Launch Site Names

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

Landing Outcome Cont.

```
[30]: %sql ibm_db_sa://yyy33800:dwNKg8J3L0IBd6CP@1bbf73c5
      %sql SELECT Unique(LAUNCH_SITE) FROM SPACEXTBL;

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

[30]:

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

Records with Launch Site Starting with CCA

- Displaying 5 records below

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

Total Payload Mass

Total Payload Mass

- **45,596 kg** (total) 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

Average Payload Mass

- **2,928 kg** (average) carried by booster version F9 v1.1

```
%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

Landing & Mission Info

1st Successful Landing in Ground Pad

- 12/22/2015

```
%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

Booster Drone Ship Landing

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

```
%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 Failed 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

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

Failed Landings on Drone Ship

In 2015

- 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)

Count of Successful Landings

- Count of landing outcomes between 2010-06-04 and 2017-03-20 in descending order

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

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

Landing _Outcome	count_outcomes
Success	20
No attempt	10
Success (drone ship)	8
Success (ground pad)	6
Failure (drone ship)	4
Failure	3
Controlled (ocean)	3
Failure (parachute)	2
No attempt	1

Section 3

Launch Sites Proximities Analysis

Launch Sites

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

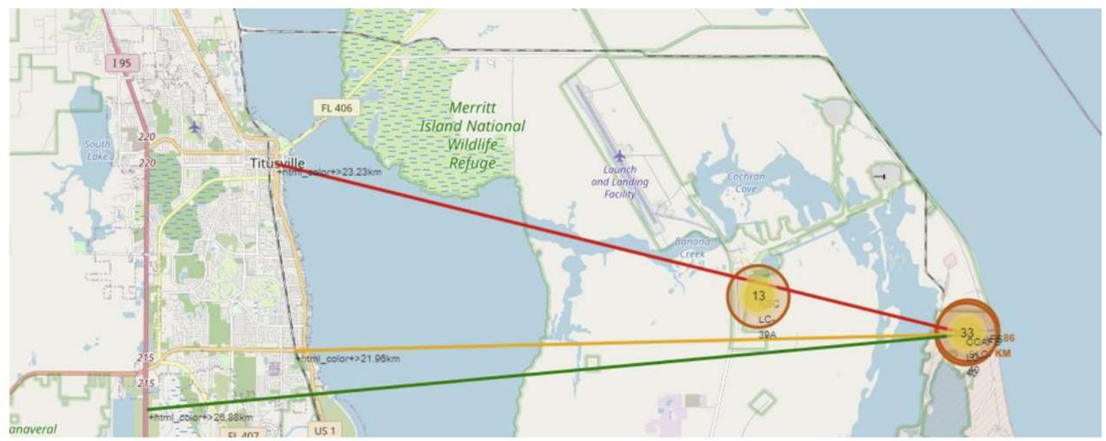
At Each Launch Site

- **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

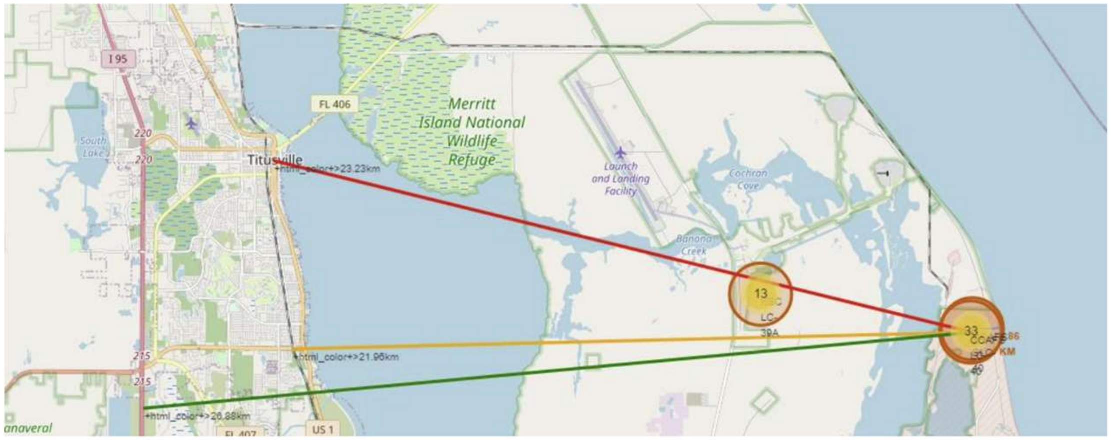
- **.86 km** from nearest coastline
- **21.96 km** from nearest railway
- **23.23 km** from nearest city
- **26.88 km** from nearest highway

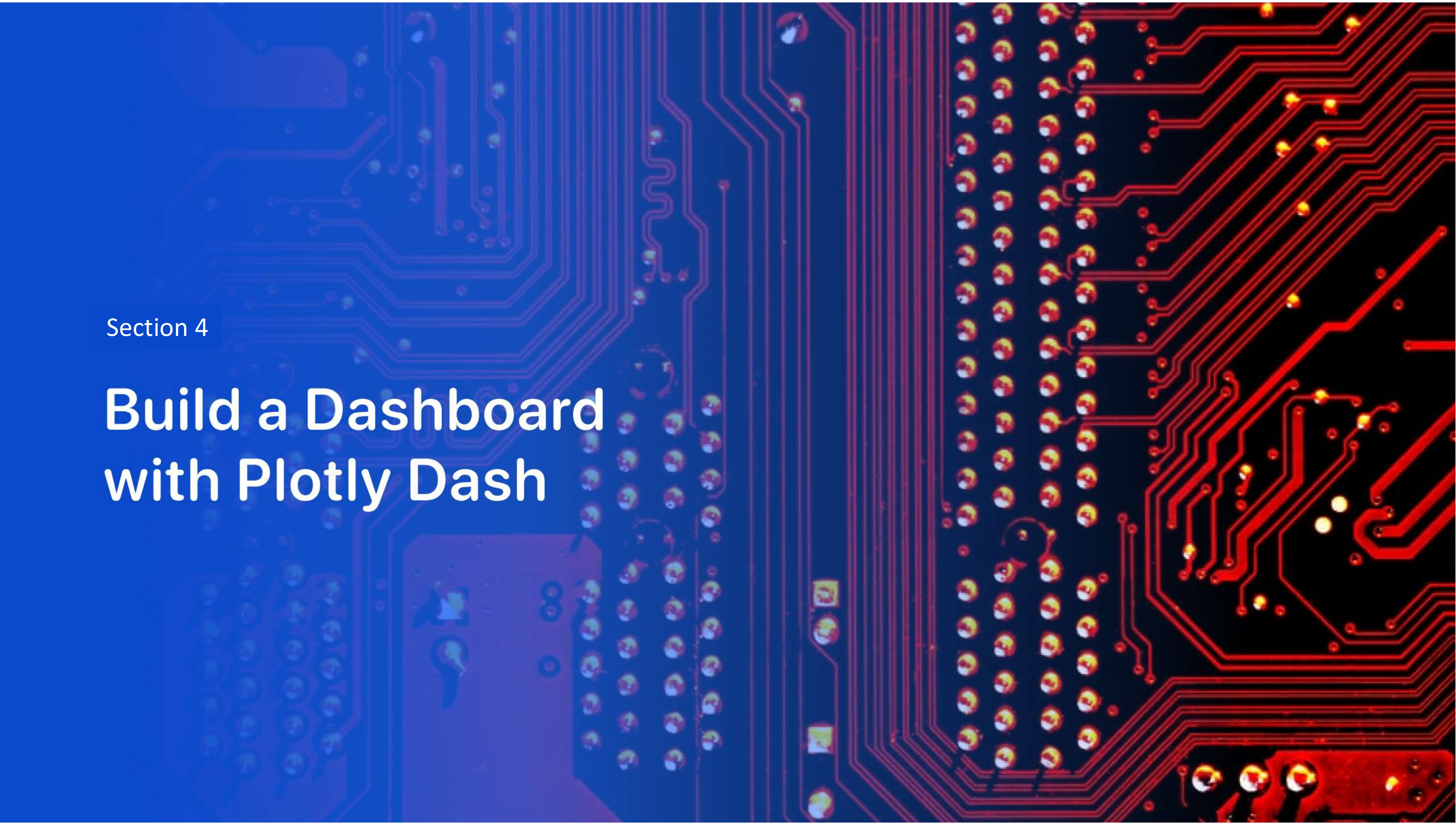


Distance to Proximities

CCAFS SLC-40

- **Coasts**: help ensure that spent stages dropped along the launch path or failed launches don't fall on people or property.
- **Safety / Security:** needs to be an exclusion zone around the launch site to keep unauthorized people away and keep people safe.
- **Transportation/Infrastructure and Cities**: need to be away from anything a failed launch can damage, but still close enough to roads/rails/docks to be able to bring people and material to or from it in support of launch activities.

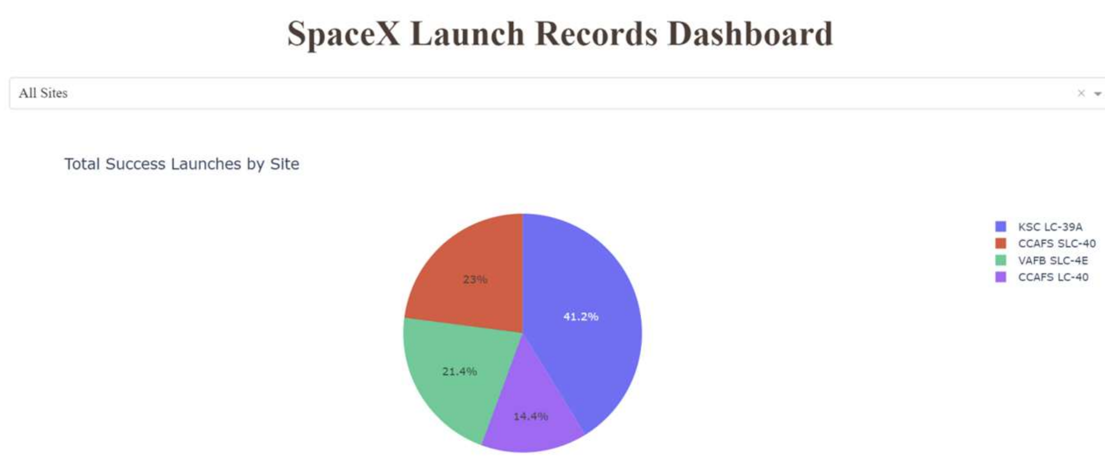



Section 4
Build a Dashboard
with Plotly Dash

Launch Success by Site

Success as Percent of Total

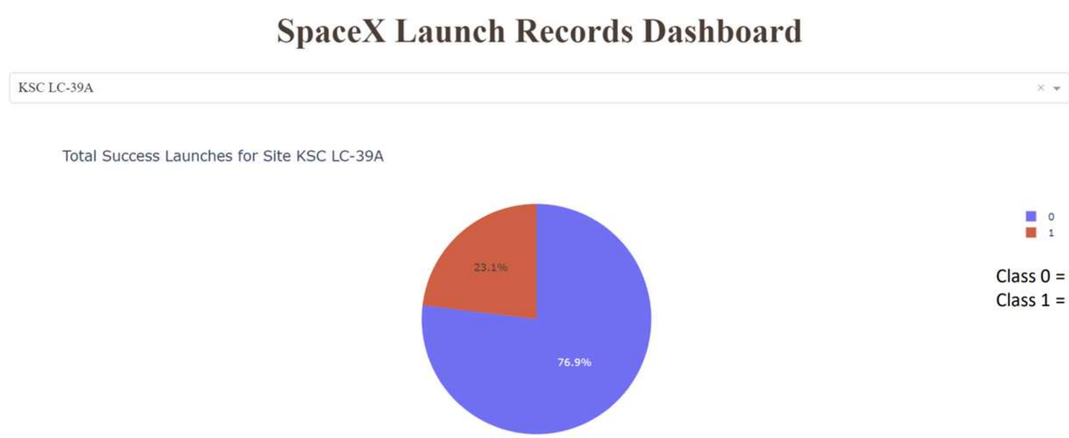
- **KSC LC-39A** has the **most successful launches** amongst launch sites (**41.2%**)



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



Class 0 = Fail
Class 1 = Success

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

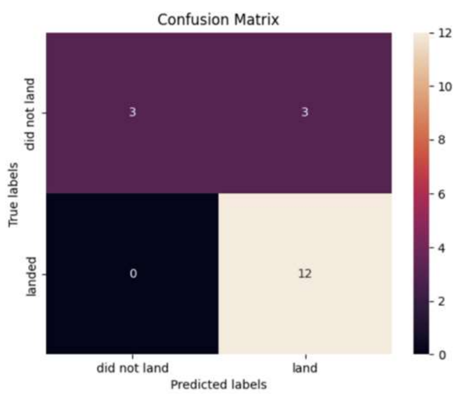
```
models = {'KNeighbors':knn_cv.best_score_,
              'DecisionTree':tree_cv.best_score_,
              'LogisticRegression':logreg_cv.best_score_,
              'SupportVector': svm_cv.best_score_}

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
- Confusion Matrix Outputs:
  - 12 True positive
  - 3 True negative
  - **3 False positive**
  - 0 False Negative



- **Precision** = TP / (TP + FP)
  - 12 / 15 = .80
- **Recall** = TP / (TP + FN)
  - 12 / 12 = 1
- **F1 Score** = 2 * (Precision * Recall) / (Precision + Recall)
  - 2 * (.8 * 1) / (.8 + 1) = .89
- **Accuracy** = (TP + TN) / (TP + TN + FP + FN) = .833

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

Conclusion

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!