Winning Space Race with Data Science

David O’Connor

17/10/2023

Outline

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

Executive Summary

- Summary of methodologies:
  - Data Collection: SQL, Web API & Web Scraping
  - Data Wrangling: Cleaning and Preparation
  - Data Visualisation
  - Exploratory Data Analysis
  - Machine Learning Methods
- Summary of all results
  - EDA
  - Analyitics
  - ML results

Introduction

- Project background and context:
    - Create a predictive model to assess the outcomes of first stage landing outcomes for Space X rocket launches
- Problems you want to find answers:
    - What method produces most accurate prediction of launch outcome

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Python Web API @ https://api.spacexdata.com/v4/rockets
  - Python Web scraping @ https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922
- Perform data wrangling
  - Check columns for percentage of missing data
  - Check data types
  - Assess number of launches
  - Create a landing class column based on successful or unsuccessful landing outcome.

Methodology

- 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 Operations: Normalize, test/train split, train and evaluate using k-means, logistic regression, decision trees, and kNN

Data Collection

- Describe how data sets were collected:

Method 1: SpaceX API
  - API get request -> pd json normalize -> Data Wrangling

Method 2: Wikipedia Webscraping
  - http get request -> BeautifulSoup html parser -> Data Wrangling

Data Collection – SpaceX API

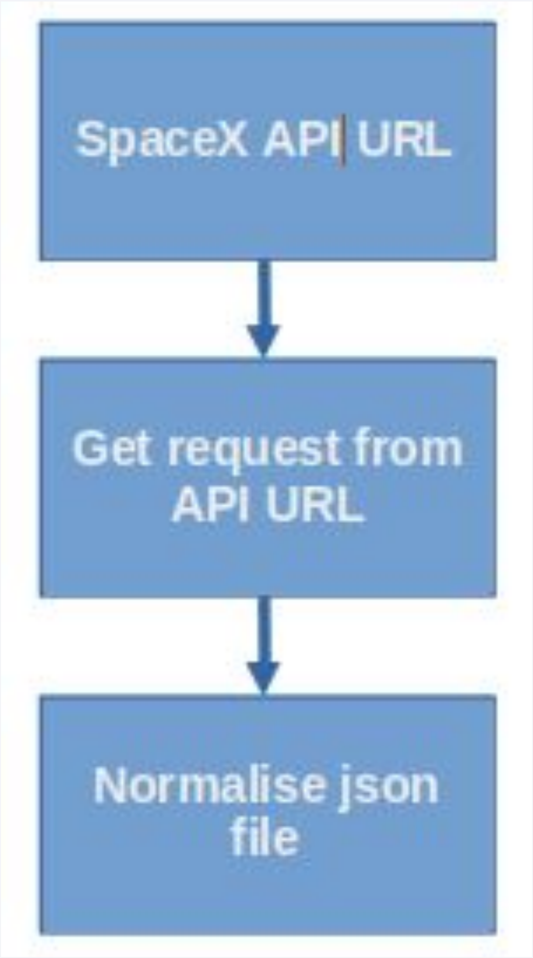
Requesting data from API

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

response = requests.get(spacex_url)

data = pd.json_normalize(response.json())

- https://github.com/skoopsy/ibm-ds-course/blob/main/a1_web_API_requests.ipynb



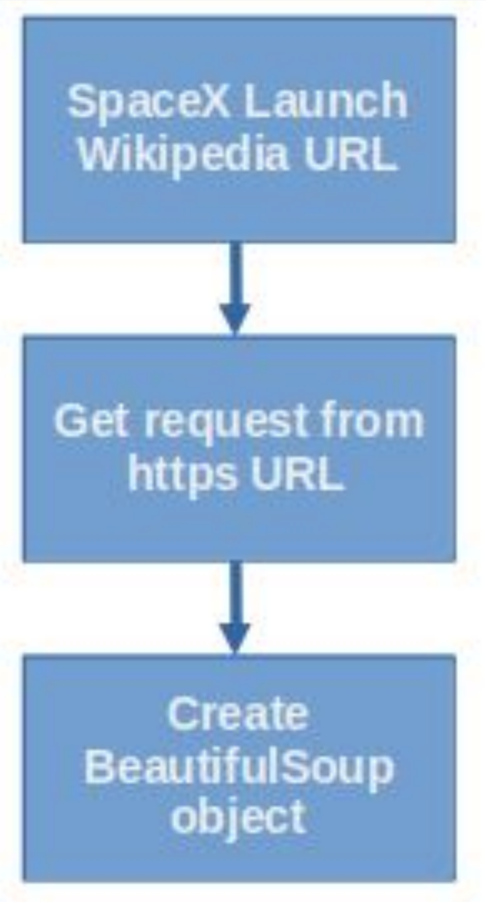
Data Collection - Scraping

static_url = "https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922"

page=requests.get(static_url)

soup=BeautifulSoup(page.text, 'html.parser')

https://github.com/skoopsy/ibm-ds-course/blob/main/a2_web_scraping_wikipedia.ipynb

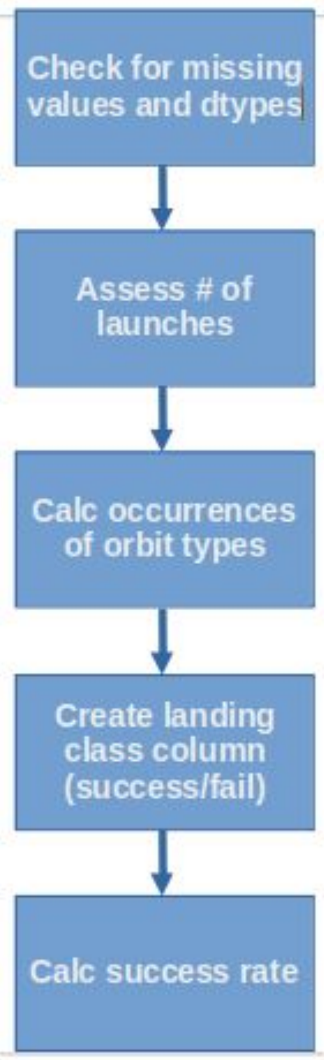


Data Wrangling

- Describe how data were processed:
    - See flow chart
    - Code:

```
df=pd.read_csv("https://cf-courses-data.s3.us.c
df.isnull().sum() / df.count()*100
df.dtypes
df.LaunchSite.value_counts()
df.Orbit.value_counts()
landing_outcomes = df.Outcome.value_counts()
landing_class = []
for i, outcome in enumerate(df.Outcome):
    if outcome in bad_outcomes:
        landing_class.append(0)
    else:
        landing_class.append(1)
df["Class"].mean().round(2)
```

- https://github.com/skoopsy/ibm-ds-course/blob/main/a3_data_wrangling-exploratory_data_analysis.ipynb



EDA with Data Visualization

Summarize what charts were plotted and why you used those charts:

- Flight Number vs Launch Site: Chronology of launch site usage
- Payload vs Launch Site: Payload capacity of each launch site
- Success rate vs Orbit type: Are certain orbit types more successful
- Flight Number vs Orbit Type: Chronology of orbit type launches
- Payload vs Orbit Type: Payload capacity of orbit types
- Success Rate vs Year: Is SpaceX becoming better at landing rockets?

• https://github.com/skoopsy/ibm-ds-course/blob/main/a5_eda_visualisation_seaborn.ipynb

EDA with SQL

- Using bullet point format, summarize the SQL queries you performed

```
df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/l
df.to_sql("SPACEXTBL", con, if_exists='replace', index=False,method="multi")
%sql create table SPACEXTABLE as select * from SPACEXTBL where Date is not null
%sql SELECT DISTINCT LAUNCH_SITE from SPACEXTABLE;
%sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where customer='NASA (CRS)';
%sql SELECT AVG(PAYLOAD_MASS__KG_) from SPACEXTBL WHERE BOOSTER_VERSION='F9 v1.1';
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE 'Success (ground pad)';
%sql SELECT BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ bet
%sql select count(MISSION_OUTCOME), mission_outcome from SPACEXTBL GROUP BY MISSION_OUTCOME like "Success%";
%sql select booster_version from SPACEXTBL where payload_mass__kg_=(select MAX(payload_mass__kg_) from SPACEXTBL)
%sql select LANDING_OUTCOME, booster_version, launch_site, date from SPACEXTBL where date like "2015%" and LANDIN
%sql select LANDING_OUTCOME, count(*) from SPACEXTBL where date between "2010-06-04" and "2017-03-20"
```

- https://github.com/skoopsy/ibm-ds-course/blob/main/a4_sql_exploratory_data_analysis.ipynb

Build an Interactive Map with Folium

- Summarize what map objects such as markers, circles, lines, etc. you created and added to a folium map
  - Mark all launch sites on map
  - use circle to mark the launch site with popup
  - create marker clusters for different zoom levels
  - successful launch = green marker
  - failed launch = red marker
- Explain why you added those objects
  - OBjects added to create a visual representation of where successful and failed launch sites occur to see if there there is any visual indication of predicting launch outcome
- https://github.com/skoopsy/ibm-ds-course/blob/main/a6_folium.ipynb

Build a Dashboard with Plotly Dash

- Summarize what plots/graphs and interactions you have added to a dashboard
  - Pie chart total successful launches
  - Pie chart for failed vs successful launches for specific launch site
  - Slider to select payload ranges
  - Payload vs Launch success
- Explain why you added those plots and interactions
  - To uncover if payload and launch site impact launch success
- https://github.com/skoopsy/ibm-ds-course/blob/main/a7_plotly_dash.py

Predictive Analysis (Classification)

- Summarize how you built, evaluated, improved, and found the best performing classification model

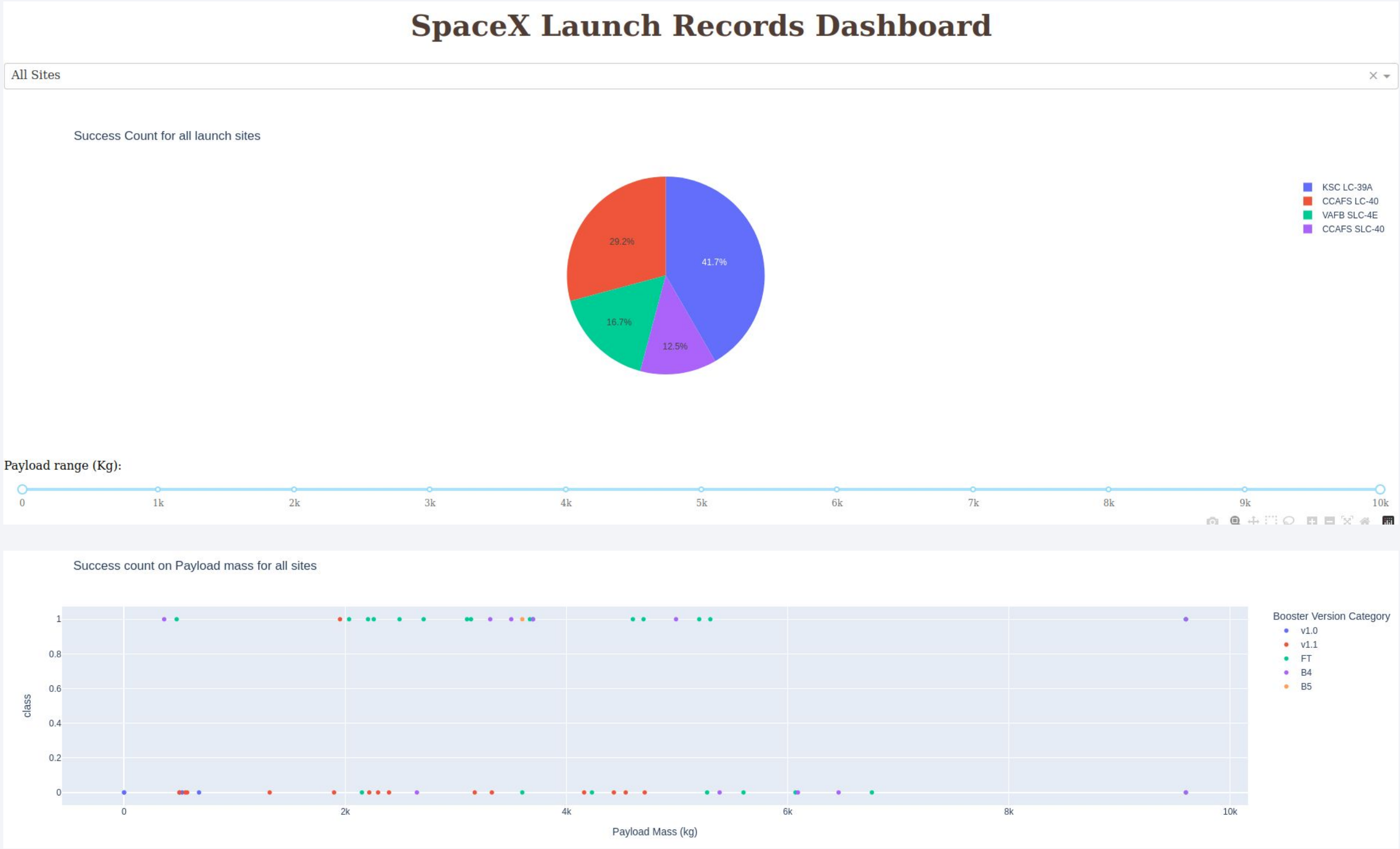
- The best model was found by comparing accuracy scores and looking for MAX
- https://github.com/skoopsy/ibm-ds-course/blob/main/a8_ml.ipynb

Results

- Exploratory data analysis results:
  - 4 launch sites
  - total payload = 45596 kg
  - avg payload = 2928.4kg
  - first successful landing outcome in 2015
  - landing success rate improved over time

Results

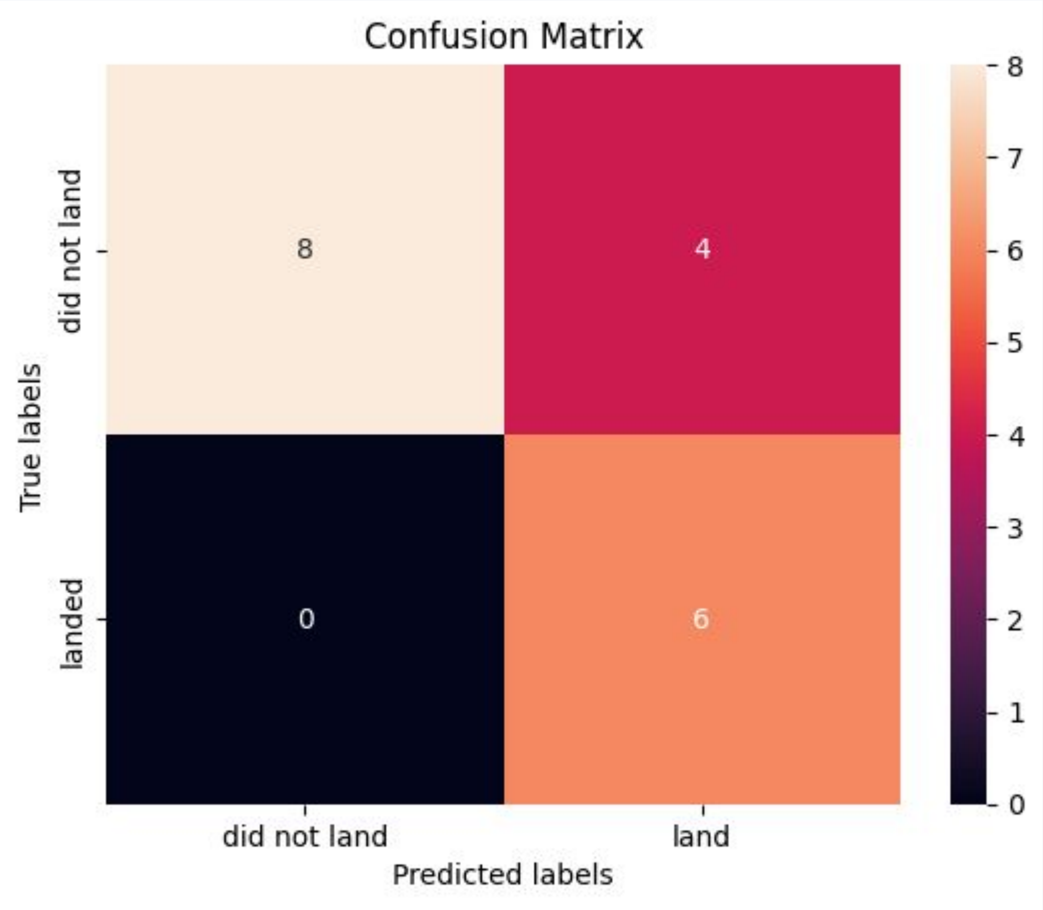
- Interactive analytics demo in screenshots (Plotly Dash):



Results

- Predictive analysis results:

Log Reg

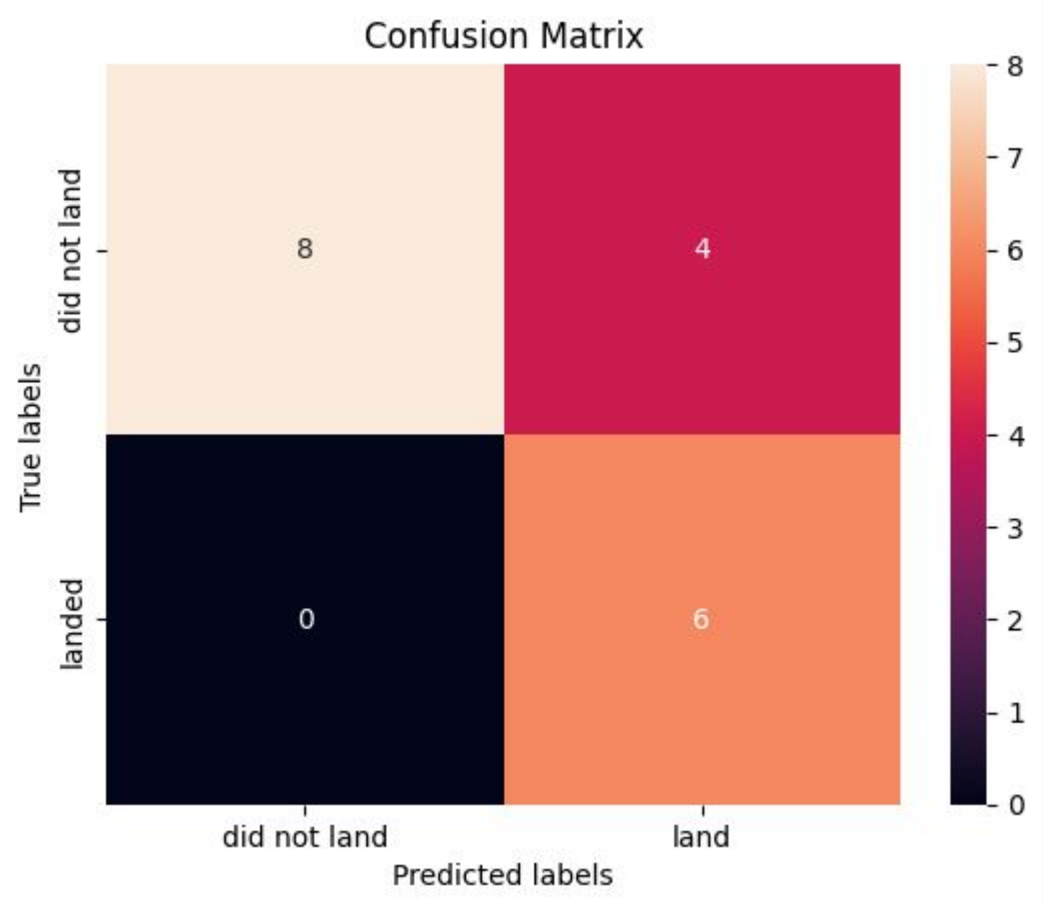


SVM

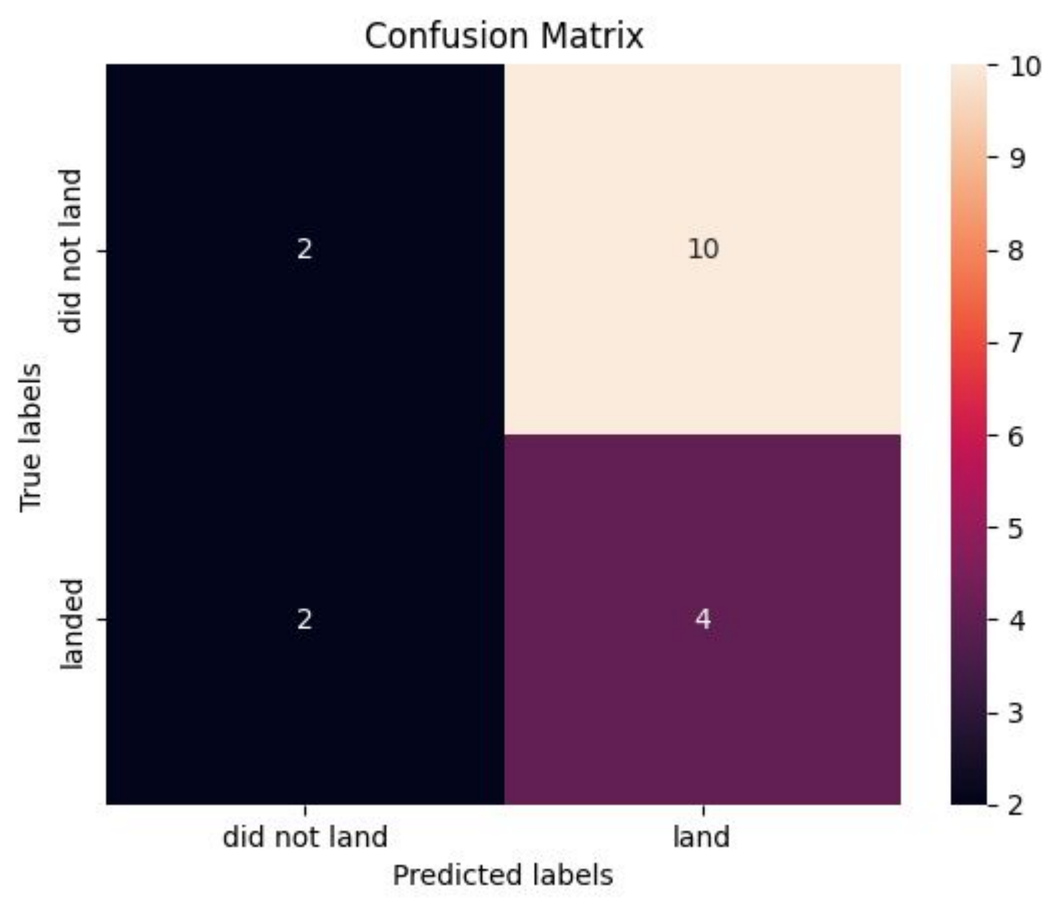
Results

- Predictive analysis results:

Tree



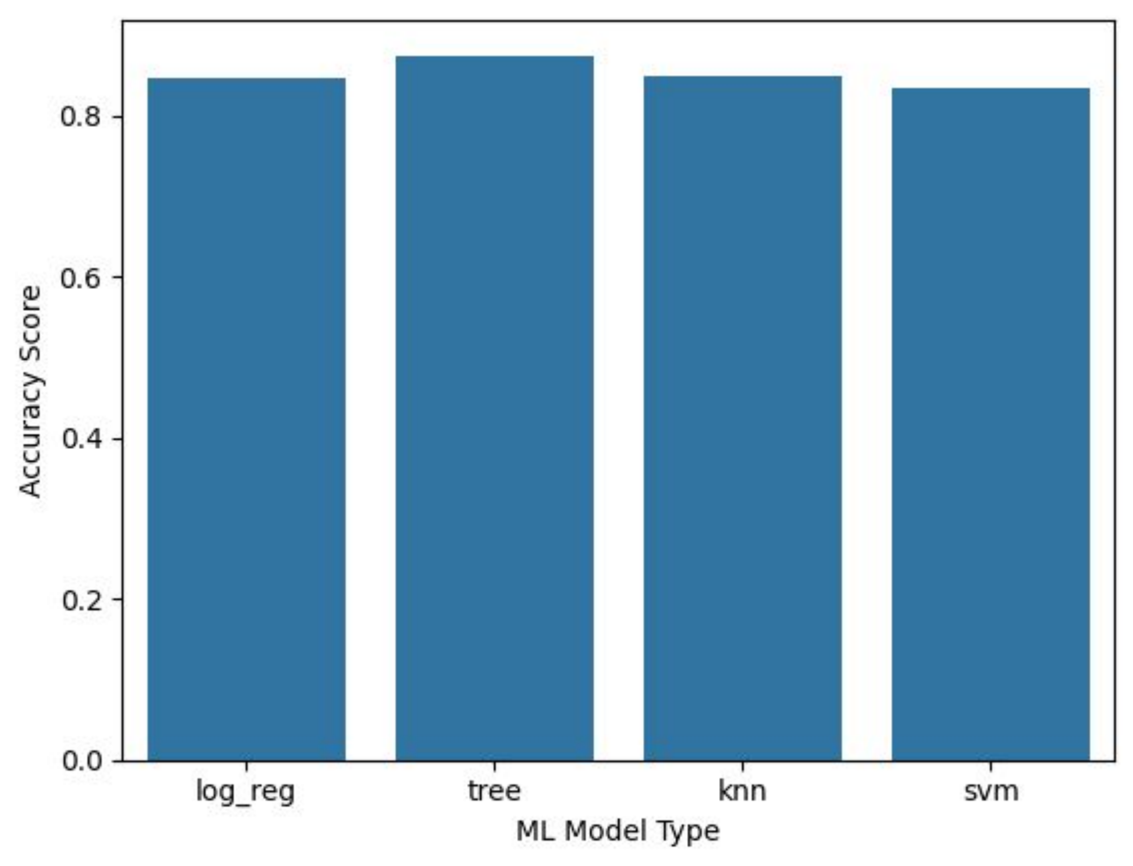
KNN



Results

- Predictive analysis results:

Accuracy Score comparison

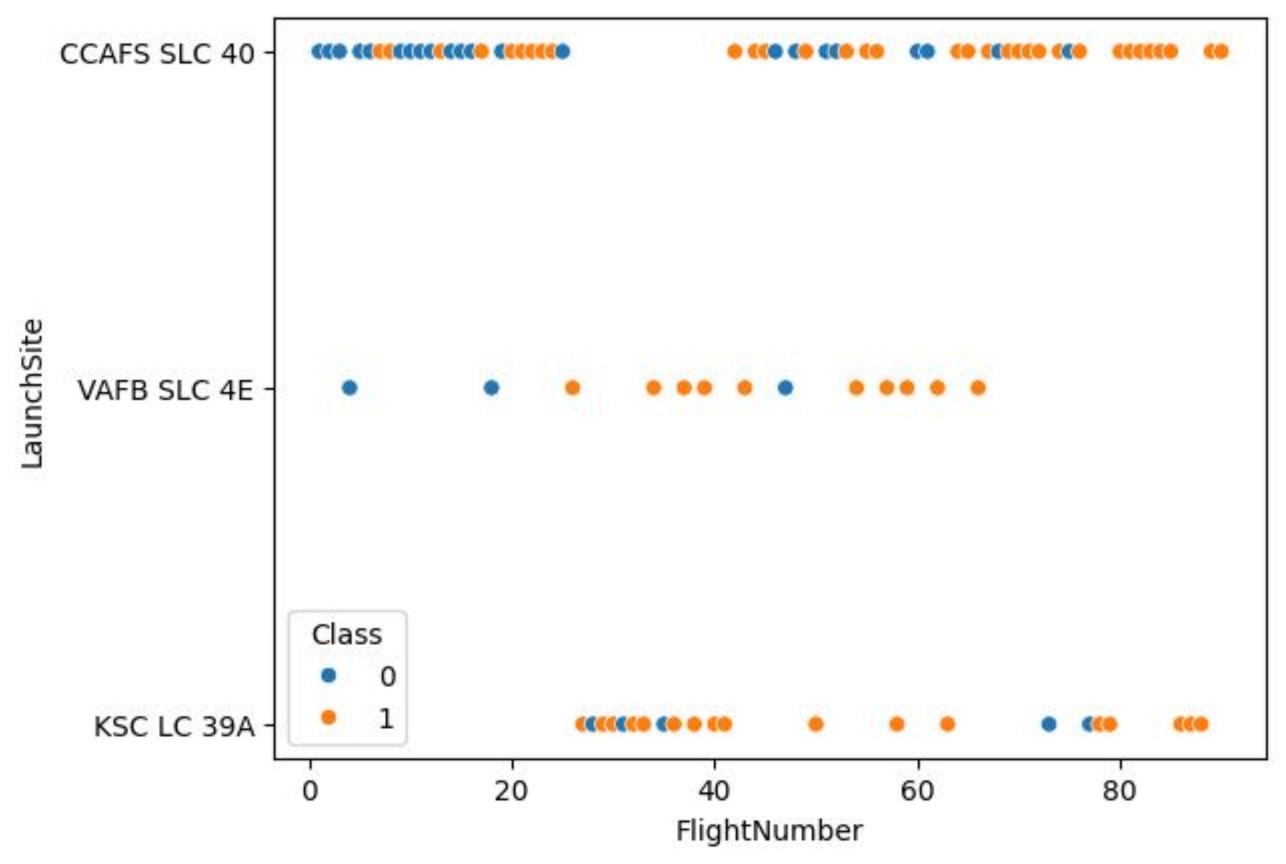


Section 2

Insights drawn from EDA

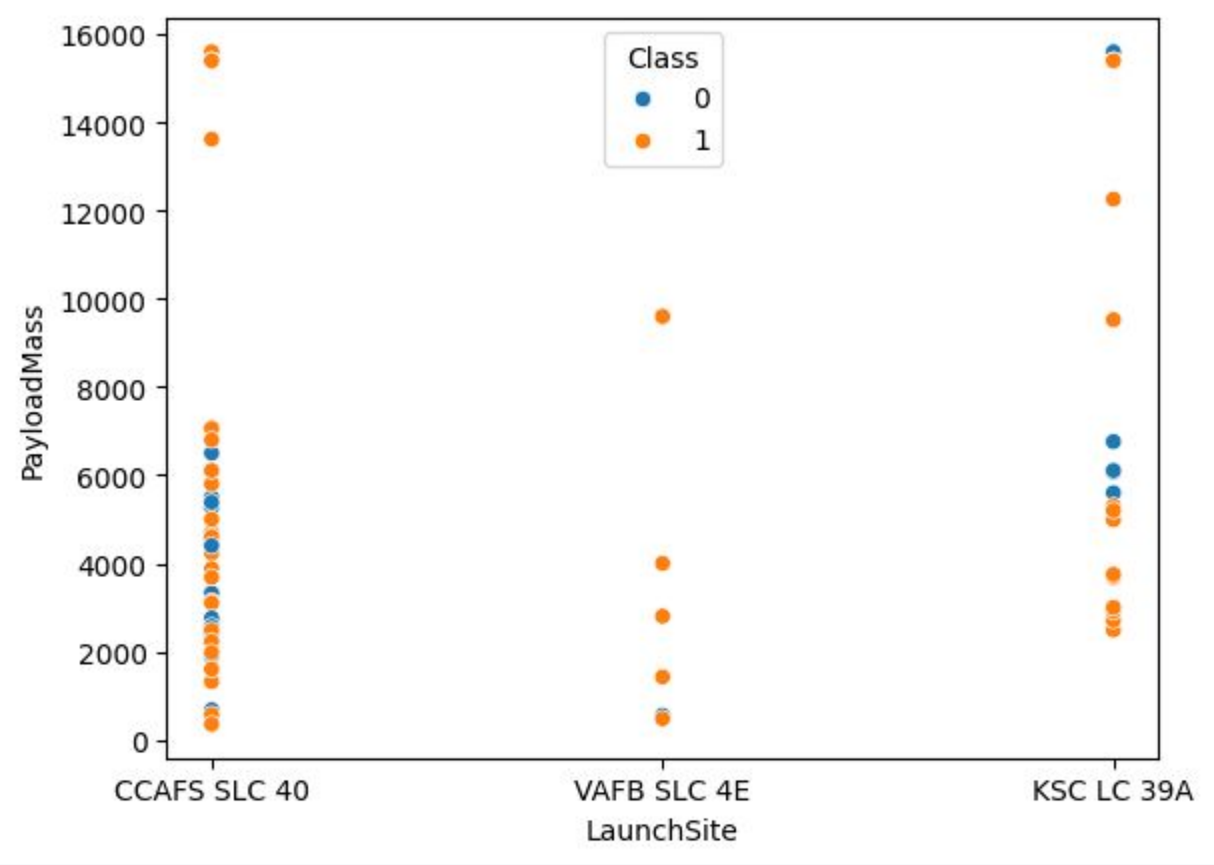
Flight Number vs. Launch Site

- higher flight numbers are more successful
- unclear if due to launch site or improvement of company processes



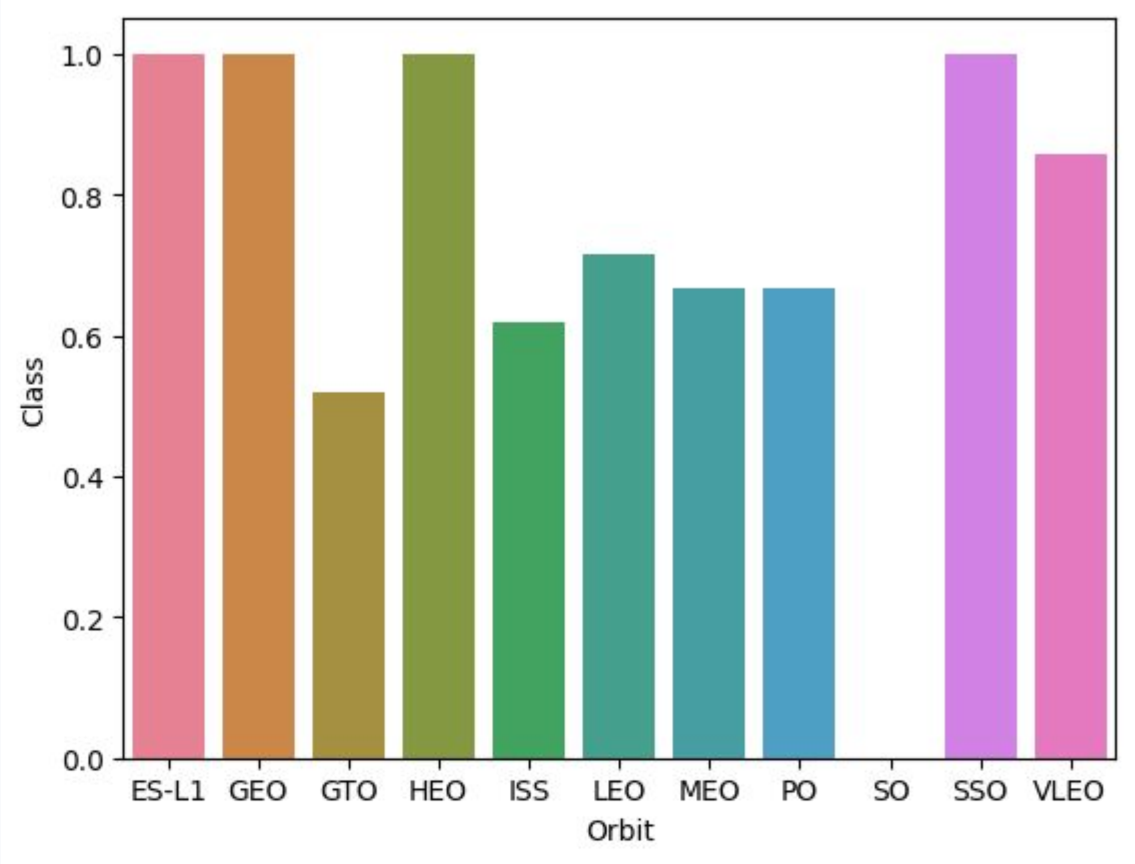
Payload vs. Launch Site

- Launch site VAFB SLC 4E has a high success rate across all payload sizes
- CCAFS SLC 40 has no failures at high payloads
- ksc lc 39a shows failures across the full payload range



Success Rate vs. Orbit Type

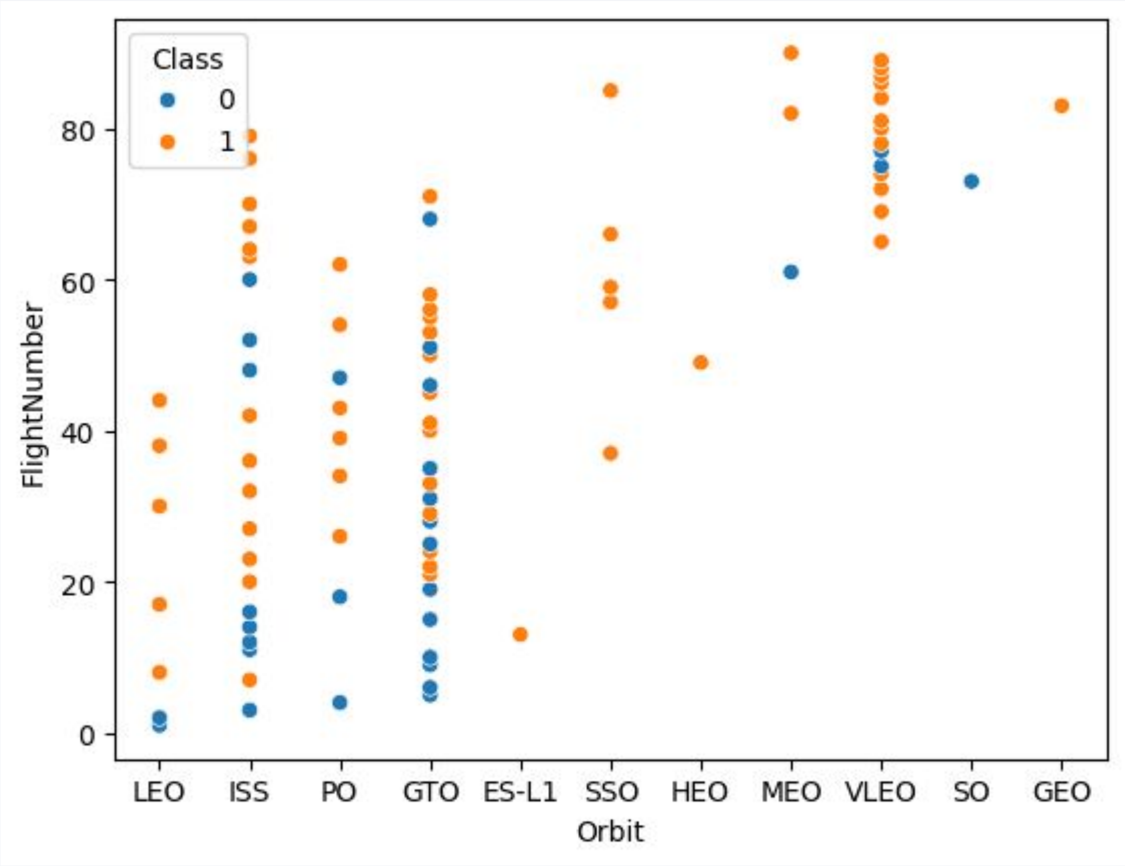
- Orbit types: ES-L1, GEO, HEO, SSO have a 100% success rate



Flight Number vs. Orbit Type

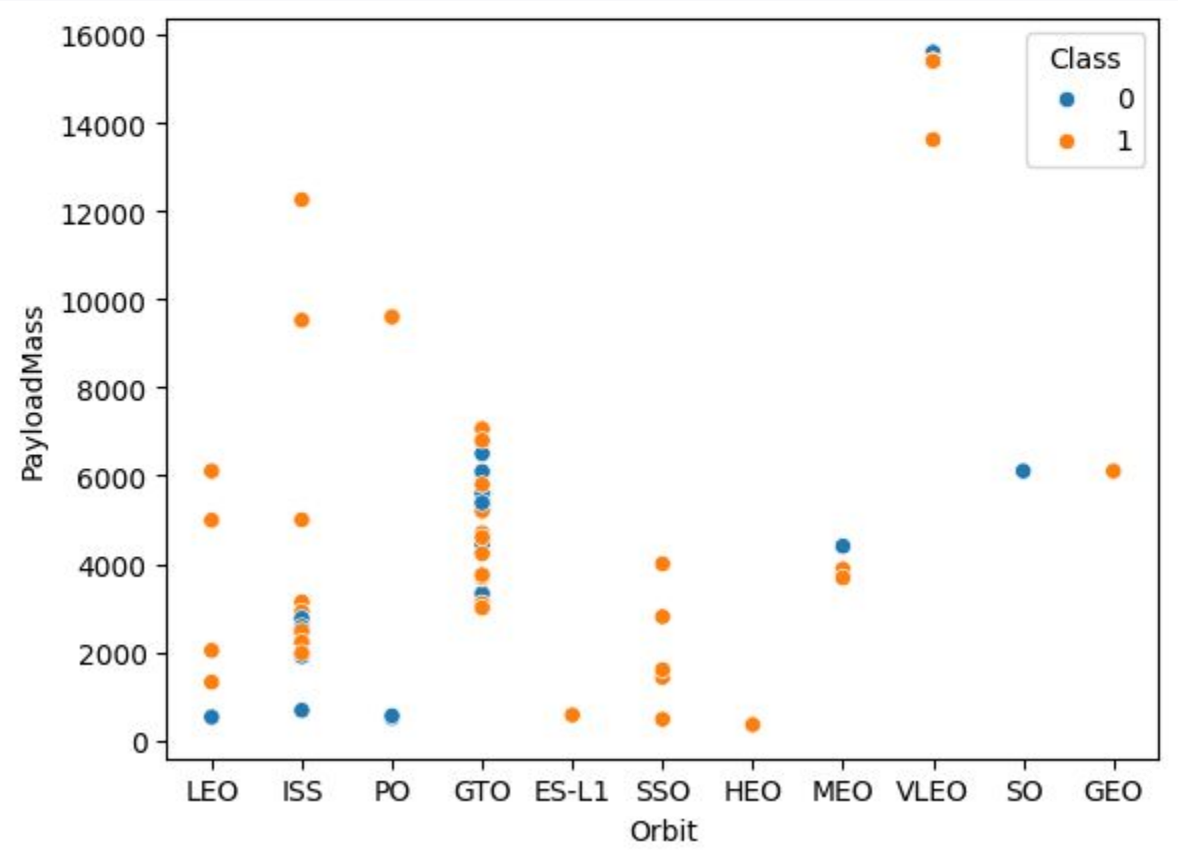
Orbits SSO to GEO were later in the launch chronology.

Less failures at higher flight numbers



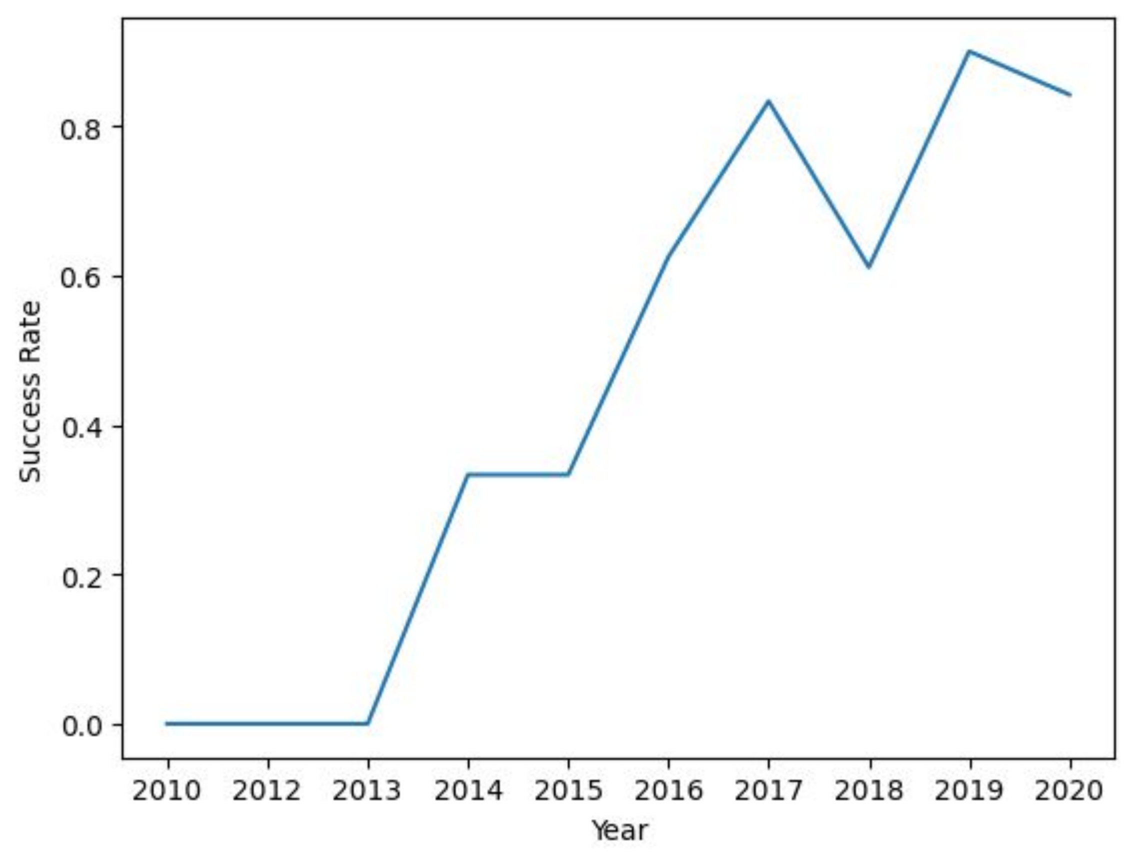
Payload vs. Orbit Type

- ISS orbit has a large range of pay loads and a high success rate (class 1)
- Launch sites SO and GEO do not have many launches



Launch Success Yearly Trend

- first 3 years were not successful
- success rate has mostly increased since 2013



All Launch Site Names

- Names of the unique launch sites:

```
%sql SELECT DISTINCT LAUNCH_SITE from SPACEXTABLE;
```

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

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` (Cape Canaveral)
- ```
  %sql SELECT LAUNCH_SITE from SPACEXTBL where (LAUNCH_SITE) LIKE 'CCA%' LIMIT 5;
  ```

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

Total Payload Mass

- Total payload carried by boosters from NASA:
-

```
%sql select sum(PAYLOAD_MASS__KG_) from SPACEXTBL where customer='NASA (CRS)';
```

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1:

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

AVG(PAYLOAD_MASS__KG_)
2928.4

First Successful Ground Landing Date

- Dates of the first successful landing outcome on ground pad:
-
```
%sql SELECT MIN(DATE) FROM SPACEXTBL WHERE LANDING_OUTCOME LIKE 'Success (ground pad)';
```

MIN(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

```
%sql SELECT BOOSTER_VERSION from SPACEXTBL where LANDING_OUTCOME='Success (drone ship)' and PAYLOAD_MASS__KG_ between
```

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

-

Total Number of Successful and Failure Mission Outcomes

- Number of successful and failure mission outcomes:
-

```
%sql select count(MISSION_OUTCOME), mission_outcome from SPACEXTBL GROUP BY MISSION_OUTCOME like "Success%";
```

count(MISSION_OUTCOME)	Mission_Outcome
1	Failure (in flight)
100	Success

Boosters Carried Maximum Payload

- Maximum payload boosters
- ```
  %sql select 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
  ```

2015 Launch Records

- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

```
%sql select LANDING_OUTCOME, booster_version, launch_site, date from SPACEXTBL where date like "2015%" and
LANDING_OUTCOME="Failure (drone ship)"
```

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

- 2 failures in time frame

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

- Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

```
%sql select LANDING_OUTCOME, count(*) from SPACEXTBL where date between "2010-06-04" and "2017-03-20"
```

Outcome	Count
NO ATTEMPT	10
Failure (Drone)	5
Success (Drone)	5
Controlled (Ocean)	3
Success (Ground Pad)	3
Failure (Parachute)	2
Uncontrolled (Ocean)	2
Precluded (Drone)	1

Section 3

Launch Sites Proximities Analysis

Folium: USA Launch Site Overview

- Zoomed out folium map shows 2 launch sites across the us in California and Florida.

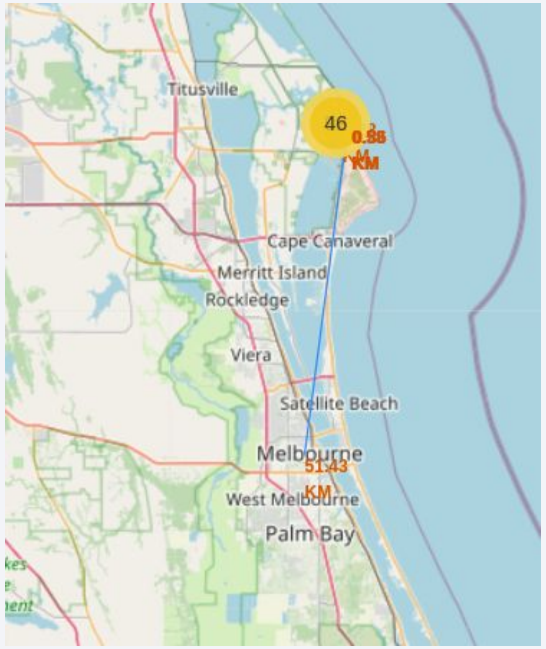
- The yellow circle indicates a mixture of failed and successful launches

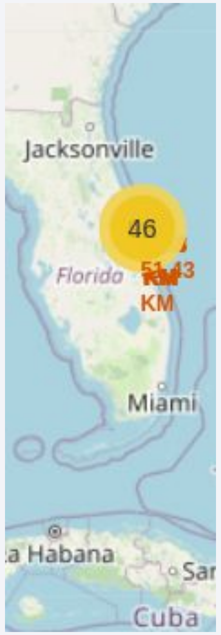
Folium: California Launch Site

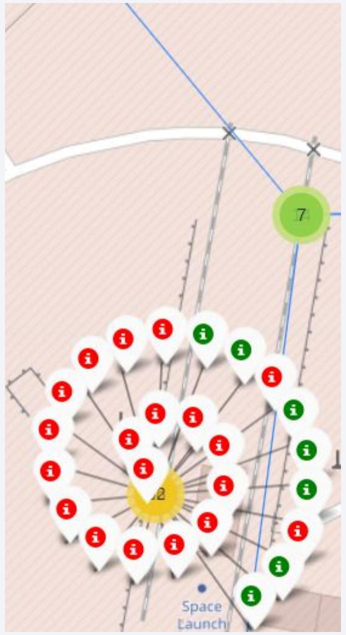
- More failed (red) than successful (green) launches can be seen at this site when zoomed in

Folium: Florida Launch Site

- Images show progressively increased magnification
- Lines depict distances from major infrastructure



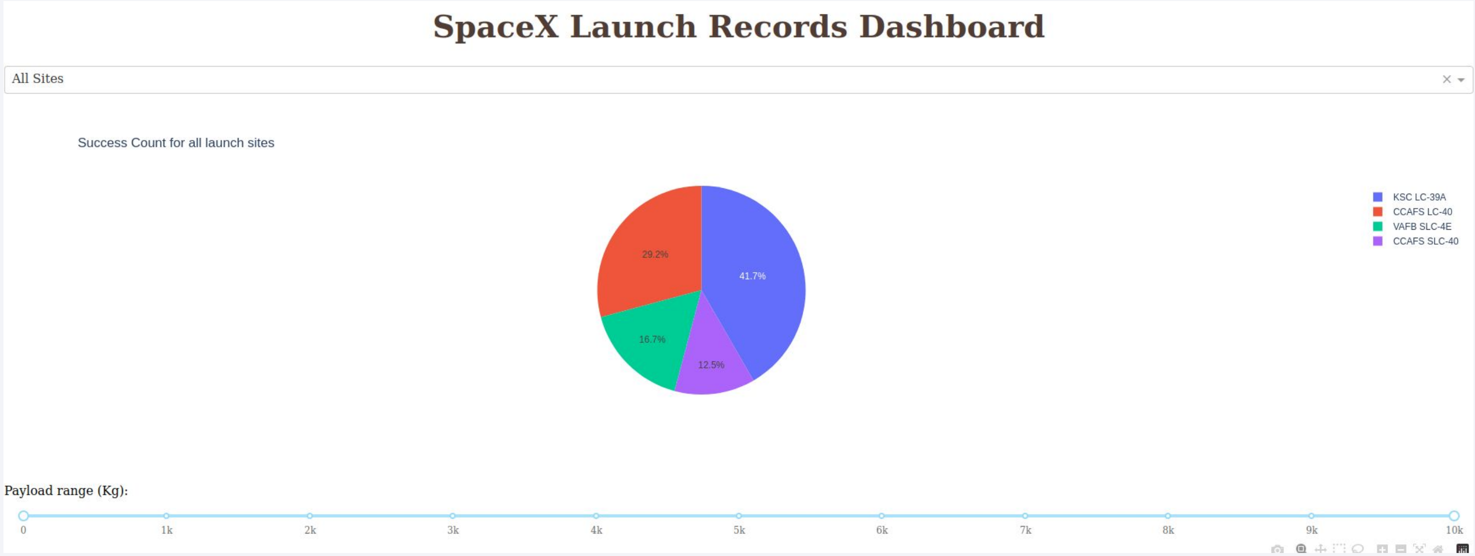




Section 4

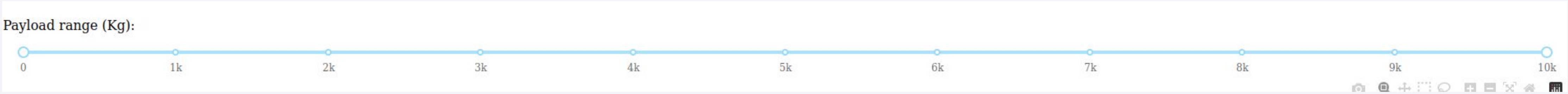
Build a Dashboard with Plotly Dash

Dash: Success Count Pie Chart



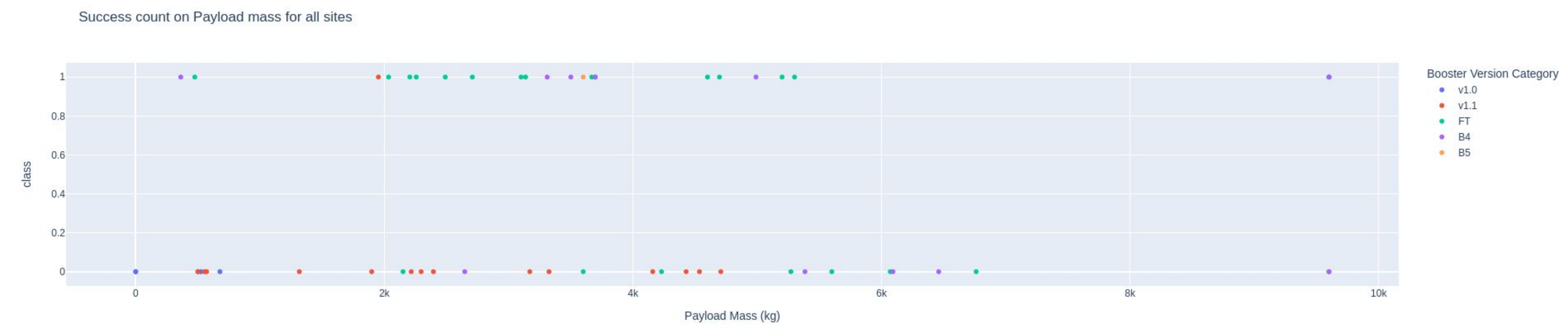
- Explain the important elements and findings on the screenshot

Dash: Payload slider



- A payload slider was created to give insights into how payload was correlated with success rate

Dash: Payload vs Launch Outcome



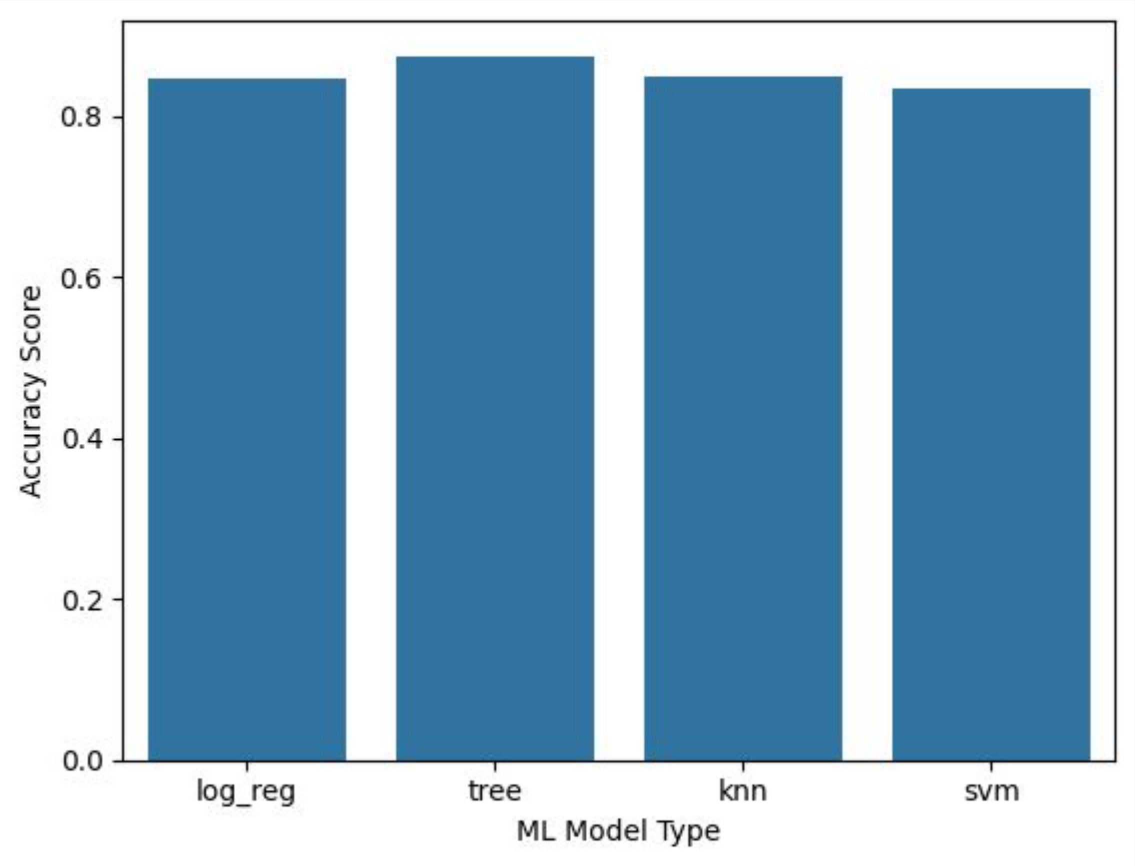
- The FT booster (Green) shows a prominent success rate at more payloads under 6k than any other booster

Section 5

Predictive Analysis (Classification)

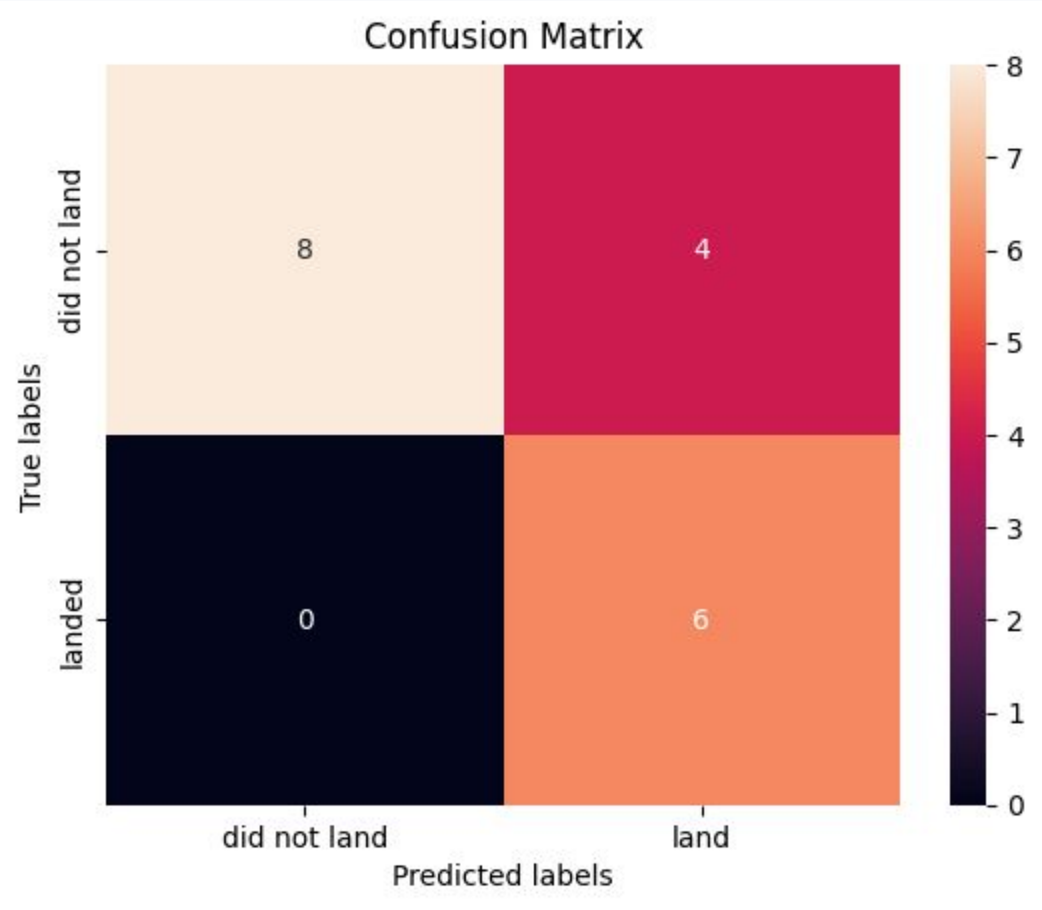
Classification Accuracy

- Comparing accuracy scores from each model via sklearn .score() on the xy test split data, Decision Tree Classifier produces the highest accuracy score
- Tree is highest with ~87%



Confusion Matrix

- Decision tree classifier confusion matrix revealing 0 false positives and 4 false negatives.



Conclusions

- Most successful launch site is KSC LC-39A
- Higher payload launches have less failures
- Successful landings have increased over time
- A decision tree classifier showed the strongest metrics for predicting landing outcome.
- …

Thank you!