IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

<Name>
<Date>

Outline

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

Executive Summary

- Summary of methodologies

Data collection from SpaceX API and SpaceX Wikipedia launch tables. Exploratory data analysis using visualisations and SQL queries. Creation of interactive visual analyses using Folium and Plotly Dash. Built different classification models to predict launch outcomes.

- Summary of all results

When predicting the launch outcome, a decision tree shows the most promising classification accuracy. As an alternative venture, it can improve launch success rates in the GTO and other orbits where SpaceX is not as developed. The FT and B4 versions of the boosters are showing promising results for payload masses of between 2,000 and 5,000 kg. Improving the launch success rate for an even higher mass and/or demonstrating even better reliability in the above-mentioned mass range would also be a way of competing with SpaceX.

Introduction

- SpaceX has announced on its website that it will be launching Falcon 9 rockets at a cost of $62 million. Other suppliers cost more than $165 million each. A large part of the savings is due to the fact that SpaceX can reuse the first stage. Therefore, if we can determine whether or not the first stage will land, we can determine the cost of a launch. This information can be used if another company wants to make a bid to SpaceX to launch a rocket. We will predict whether the first stage of the Falcon 9 will land successfully.
- What factors determine whether the first stage will land?
- How can another company compete with SpaceX in its quest for space travel?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

- Describe how data sets were collected.
- You need to present your data collection process use key phrases and flowcharts

Data Collection – SpaceX API

- Request and parse the SpaceX launch data using the GET request
- Filter Data to only include Falcon 9 launches
- Deal with Missing Values

GitHub: Hands-on Lab Complete the Data Collection APILab.ipynb

Data Collection - Scraping

- Request the Falcon9 Launch Wiki page from its URL
- Extract all column/variable names from the HTML table header
- Create a data frame by parsing the launch HTML tables

GitHub: Hands-on Lab Complete the Data Collection with Web Scraping lab.ipynb

Data Wrangling

- Dealing with Missing Values:

Calculate the mean for a column with missing values. Then use the mean to replace NaN values in the data with the mean you calculated (identify column with missing values, Calculate mean for column and Replace NAN with mean value)

- Create a landing outcome label from outcome column:

Using the column Outcome, create a list where the element is zero if the corresponding row in Outcome is unsuccessful otherwise, it's one. This variable will represent the classification variable that represents the outcome of each launch (outcome column, relabel outcome: Successful=1 unsuccessful=0 and save as classification variable)

EDA with Data Visualization

- Scatter plots: To show relationship between two variables
- Bar chart: To make comparisons between categorical variables
- Line plot: To show trends in data over intervals of time
- GitHub: Hands-on Lab EDA with Data Visualization.ipynb

EDA with SQL

- Select statements
- Select statements with operators (sum, avg, min, max)
- Select statements with range condition
- Select statements with string operator %
- Select statements with subqueries
- Select statements with grouped by column
- GitHub: Hands-on Lab EDA with SQL.ipynb

Build an Interactive Map with Folium

- Added markers and circles to highlight NASA Space Center and SpaceX launch sites
- Added marker cluster to add launch outcomes to the launch sites and to simplify a map containing many markers at the same coordinate
- Added mouse position to find coordinates of any points of interest
- Added polyline to draw lines between selected points
- GitHub: Hands-on Lab Interactive Map with Folium.ipynb

Build a Dashboard with Plotly Dash

- Added dropdown to choose between all sites or individual sites

- Added pie chart to show distribution of successful launches among different sites
- Added pie chart to show launch success rate if individual site is selected in dropdown
- Added Payload Mass (kg) range slider to filter payload mass
- Added scatterplot to show relationship between payload and launch success
- GitHub: Hands-on Lab Dashboard with Plotly Dash.py

Predictive Analysis (Classification)

- Perform steps outlined in the flowchart for Logistic Regression, SVM, Decision Tree and KNN models
  - Find model with highest accuracy in prediction
- GitHub: Hands-on Lab Predictive Analysis (Classification).ipynb

Results

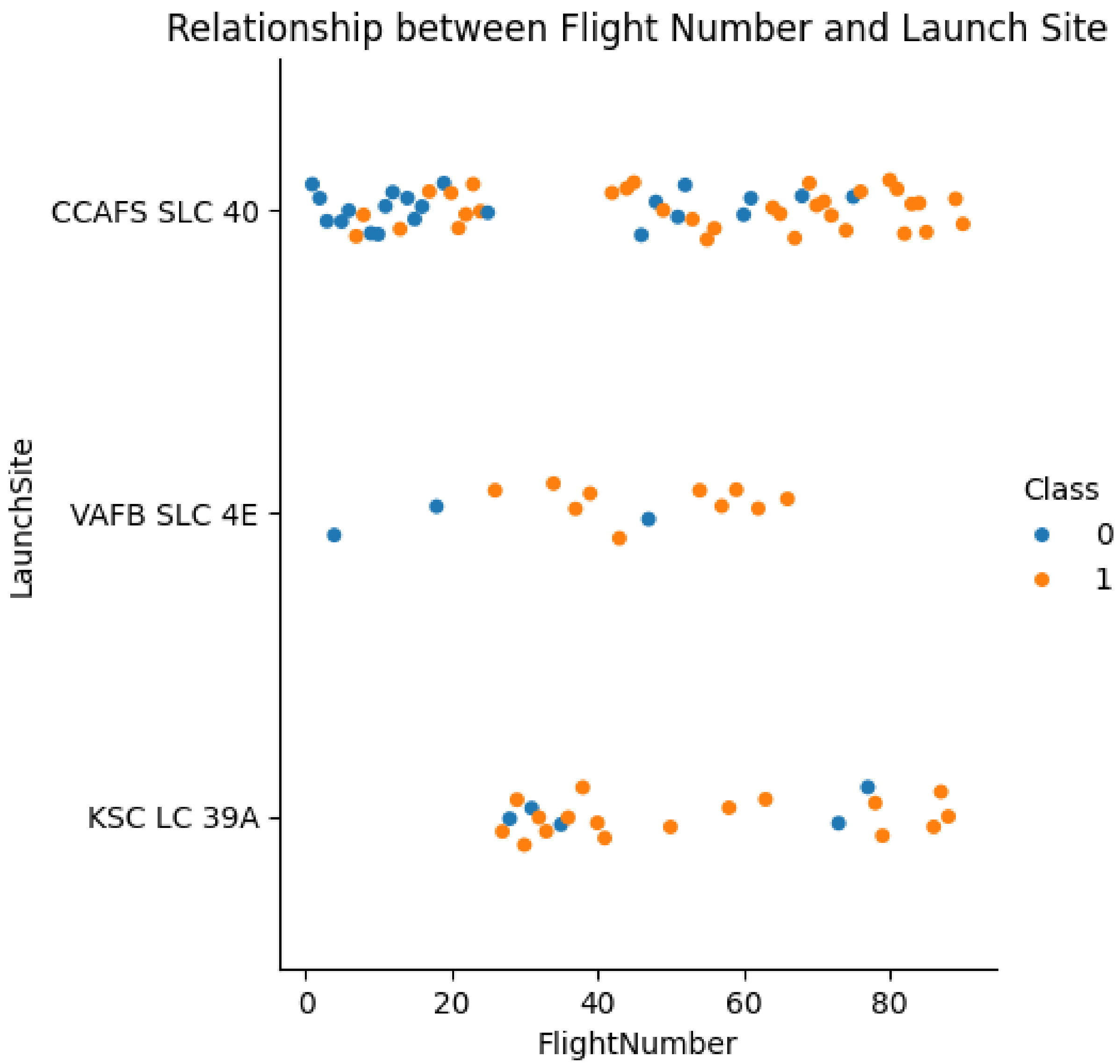
- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

Section 2

Insights drawn from EDA

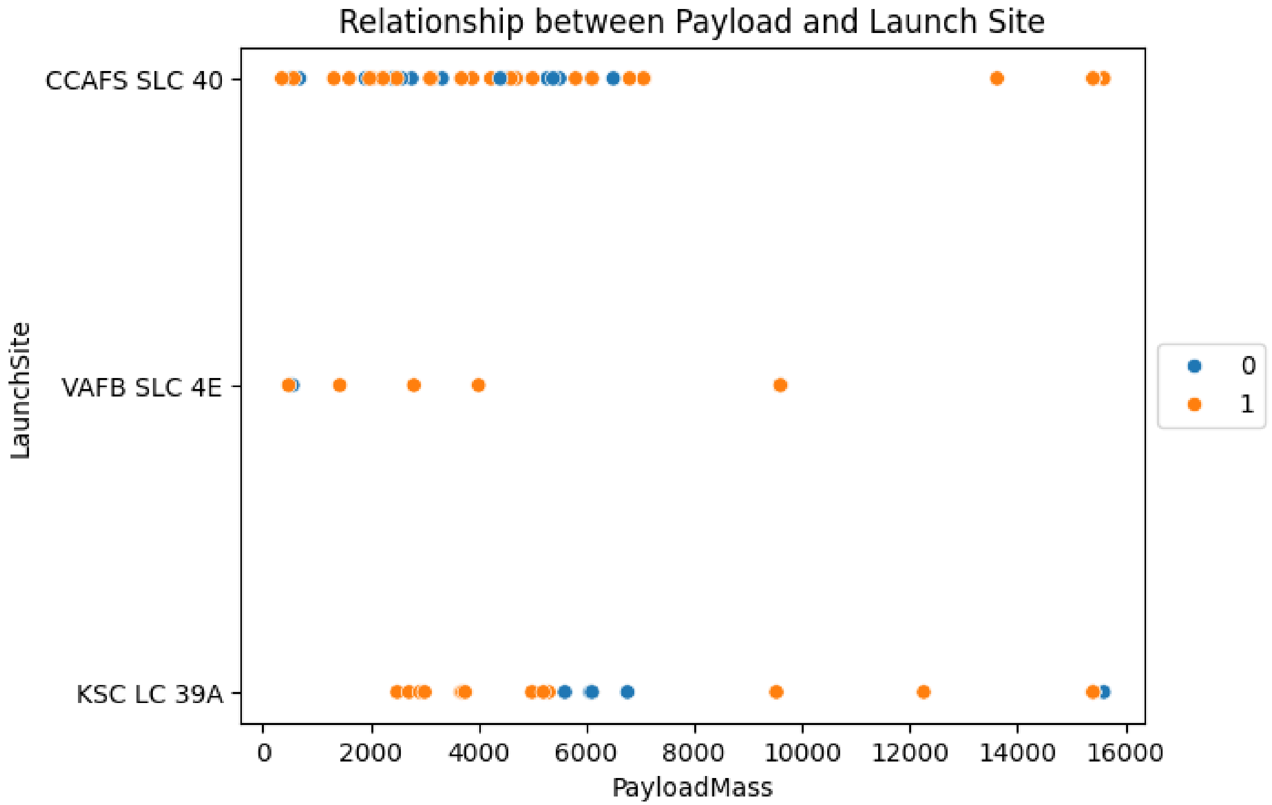
Flight Number vs. Launch Site

- CCAFS SLC 40 has the highest density of launches across various flight numbers. With increasing flight number the success rate has become better.
  • The other sites have fewer launches and less variation



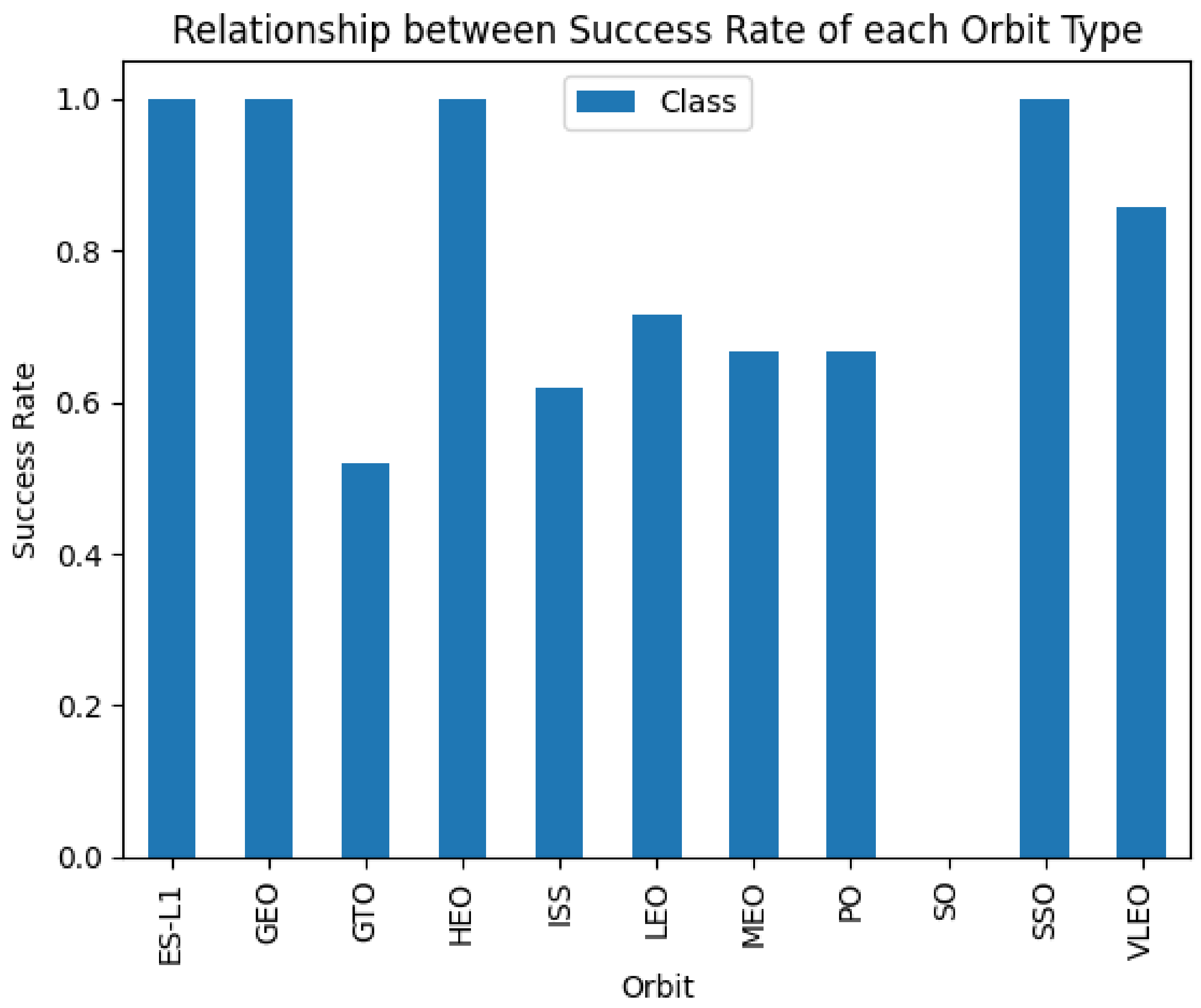
Payload vs. Launch Site

- CCAFS SLC 40 and KSC LC 39A handle a wider range of payload masses, including some of the heaviest payloads
  • VAFB SLC 4E primarily handles lighter payloads



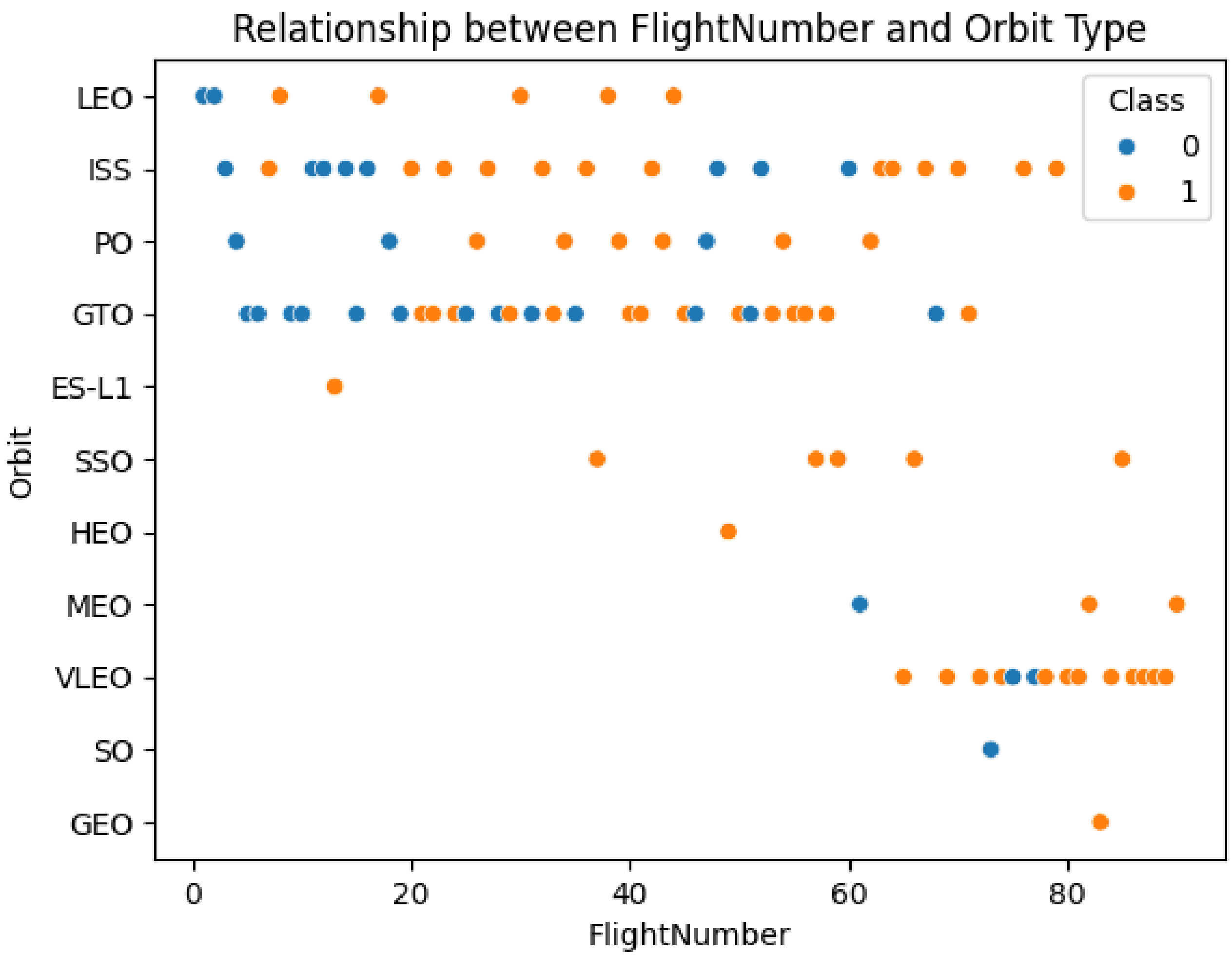
Success Rate vs. Orbit Type

- ES-L1, GEO, HEO and SSO have 100% success rate
-
- The other orbit types seem to be trickier and show varying rates



Flight Number vs. Orbit Type

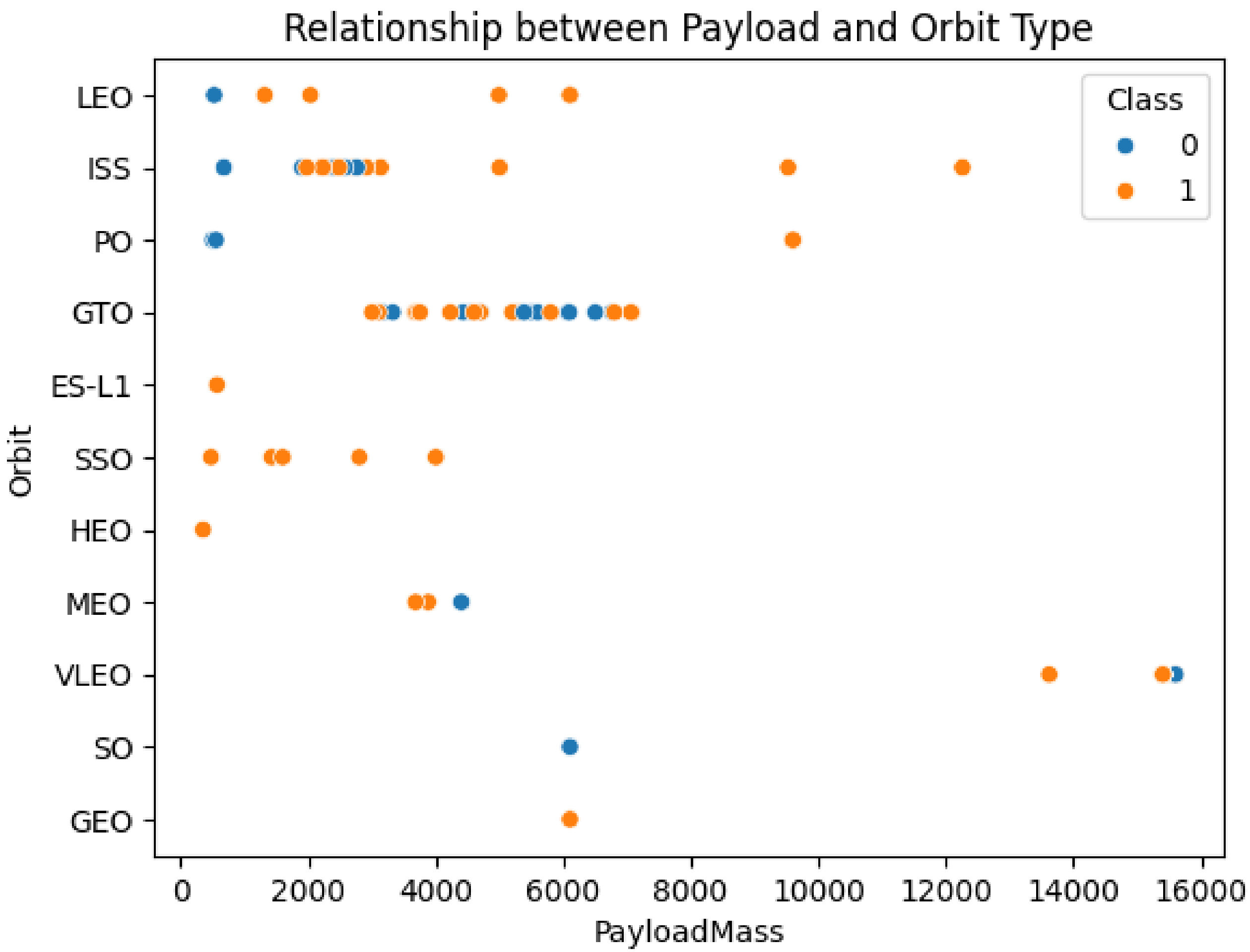
- LEO: The Success appears related to the number of flights
  • GTO: There seems to be no relationship between flight number when in GTO orbit
  • ISS, GTO and VLEO show the most number of flights



Payload vs. Orbit Type

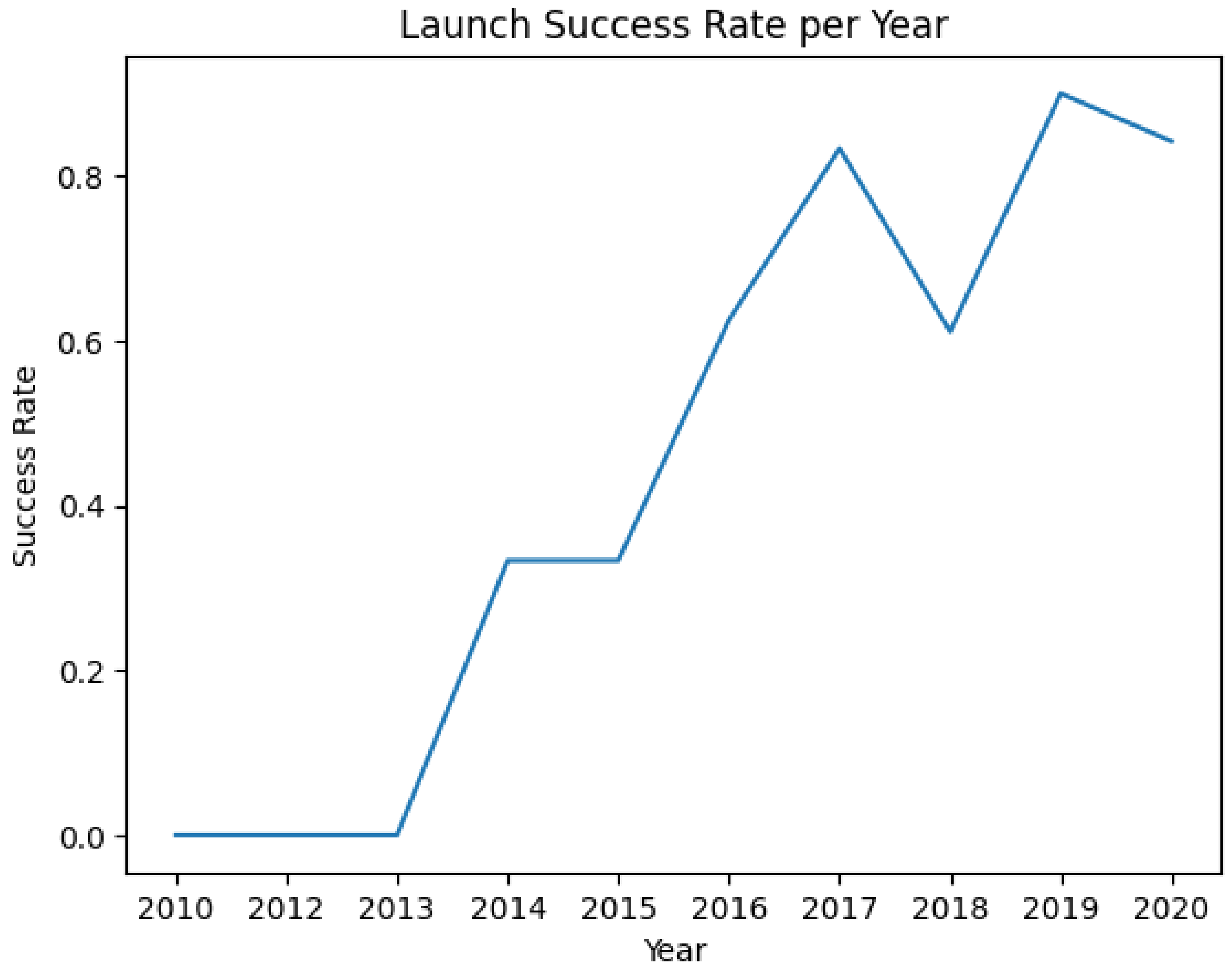
- With heavy payloads the successful landing or positive landing rate are more for Polar, LEO and ISS

- For GTO we cannot distinguish this well as both positive landing rate and negative landing (unsuccessful mission) are both there here



Launch Success Yearly Trend

- You can observe that the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing.



All Launch Site Names

- Find the names of the unique launch sites
  - %sql select distinct "Launch_Site" from SPACEXTABLE;
  - Query returned 4 unique launch sites from the data

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

Launch Site Names Begin with 'CCA'

- Find 5 records where launch sites begin with `CCA`
  • %sql select * from SPACEXTABLE where "Launch_Site" like 'CCA%' limit 5;
  • Query displays first 5 records where launch site starts with 'CCA’

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success
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
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success

Total Payload Mass

- Calculate the total payload carried by boosters from NASA
  - %sql select sum(PAYLOAD_MASS__KG_) from SPACEXTABLE where Customer = 'NASA (CRS)';
  - Shows sum of payload mass

sum(PAYLOAD_MASS__KG_)
45596

Average Payload Mass by F9 v1.1

- Calculate the average payload mass carried by booster version F9 v1.1
  - %sql select avg(PAYLOAD_MASS__KG_) from SPACEXTABLE where Booster_Version like 'F9 v1.1%';

- Shows average payload mass

avg(PAYLOAD_MASS__KG_)
2534.6666666666665

First Successful Ground Landing Date

- Find the dates of the first successful landing outcome on ground pad
  - %sql select min(DATE) from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)’;

- The first successful landing was achieved on the 22nd December in 2015

min(DATE)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

• %sql select distinct Booster_Version from SPACEXTABLE where Landing_Outcome = 'Success (drone ship)' and PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;

• Four Booster Versions have landed successfully with a payload mass greater than 4000 but less than 6000

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

Total Number of Successful and Failure Mission Outcomes

- Calculate the total number of successful and failure mission outcomes
- %sql select Mission_Outcome, count(*) as total_count from SPACEXTABLE group by Mission_Outcome;
- Count of mission outcomes

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

Boosters Carried Maximum Payload

- List the names of the booster which have carried the maximum payload mass

- %sql select Booster_Version from SPACEXTABLE where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTABLE);
- Names of different booster versions carrying maximum payload mass

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 substr(Date, 6,2) as Month, substr(Date, 0,5) as Year, Landing_Outcome, Booster_Version, Launch_Site from SPACEXTABLE where Landing_Outcome = 'Failure (drone ship)' and substr(Date, 0,5) = '2015';

- Two launches in 2015 failed to land on a drone ship

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

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(*) as Count from SPACEXTABLE where Date between '2010-06-04' and '2017-03-20' group by Landing_Outcome order by Count desc;

- Landing outcomes in descending order

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

Section 3

Launch Sites Proximities Analysis

Launch Site Location Markers

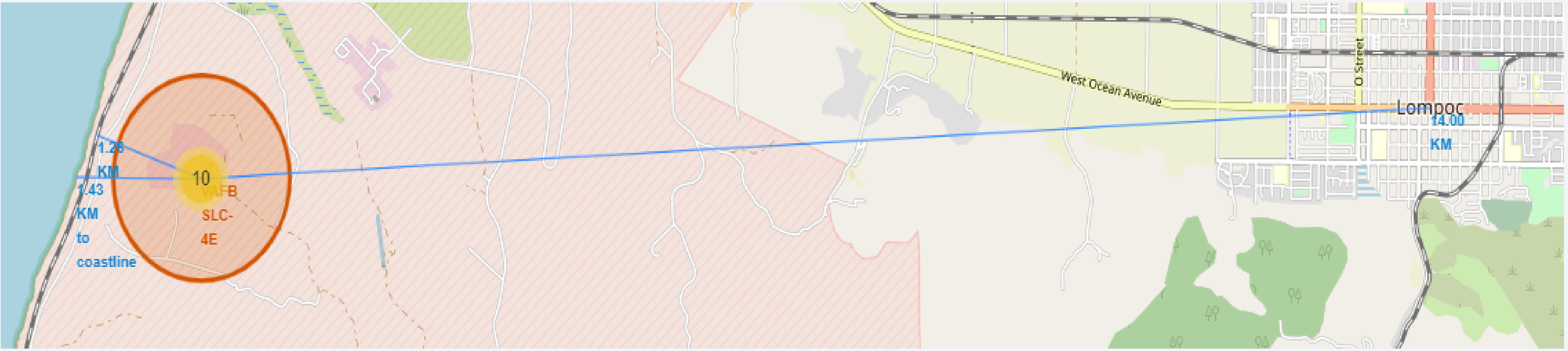
- SpaceX launch sites depicted as markers on the map
- Interestingly, launch sites are positioned close to the ocean

Launch Outcome Icons per Site

- Icons at launch site depicting outcome success in green or failure in red

Distances (km) from Launch Site

- 1.43km from nearest coastline
- 1.26km from nearest railway line
- 14.00km from nearest city (Lompoc)

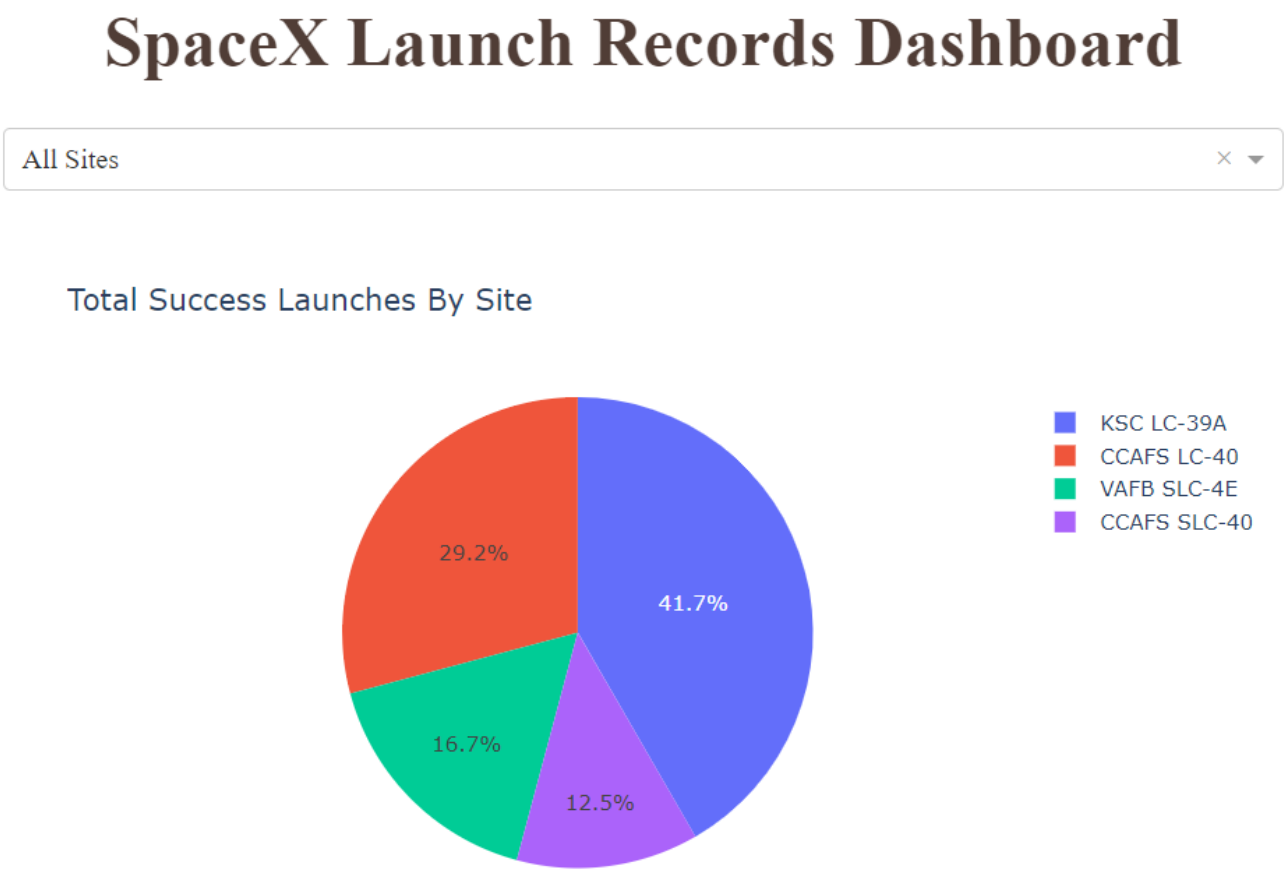


Section 4

Build a Dashboard with Plotly Dash

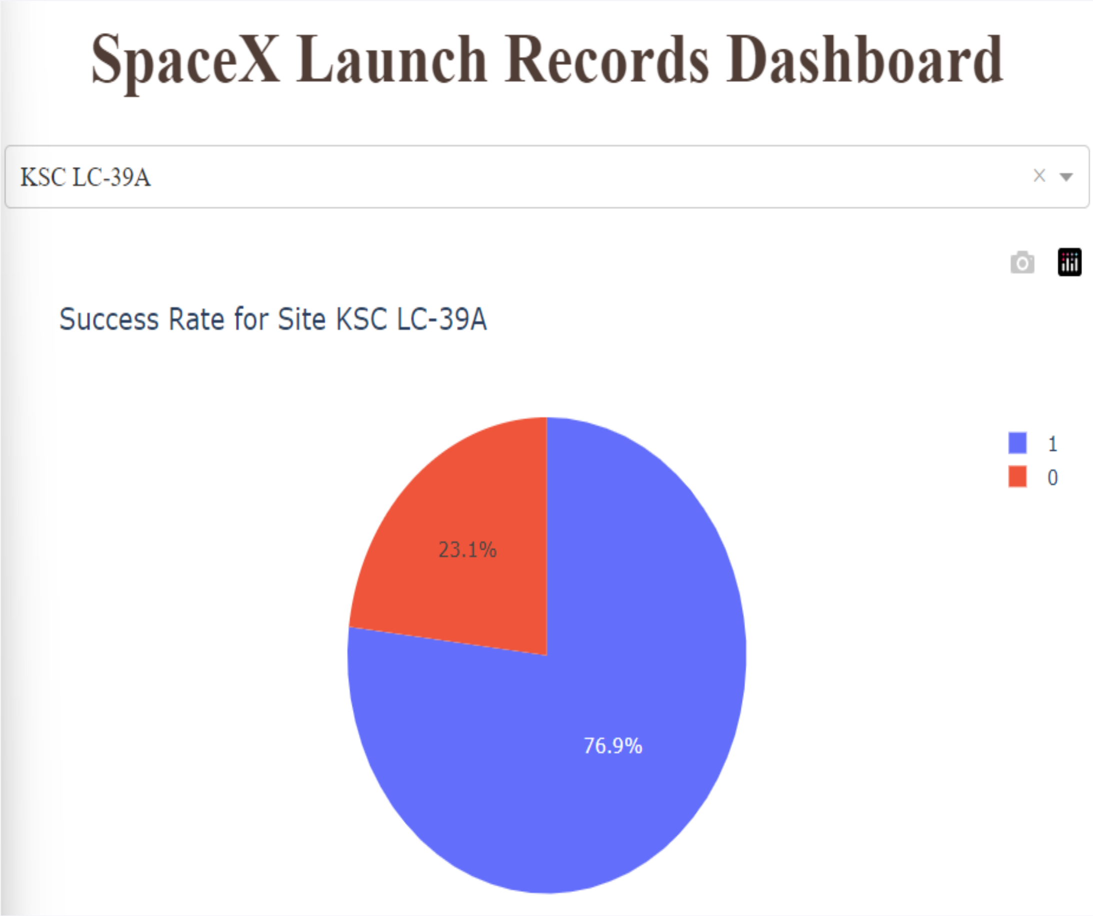
Launch Success Distribution among Sites

- Most successful launches among all launches were at sites KSC LC-39A (41.7%)
  - Followed by CCAFS LC-40 with 29.2% of overall successful launches



Launch Success Rate at KSC LC-39A

- Launch Success Rate at KSC LC-39A was 76.9%



Relationship between Payload and Success Rate

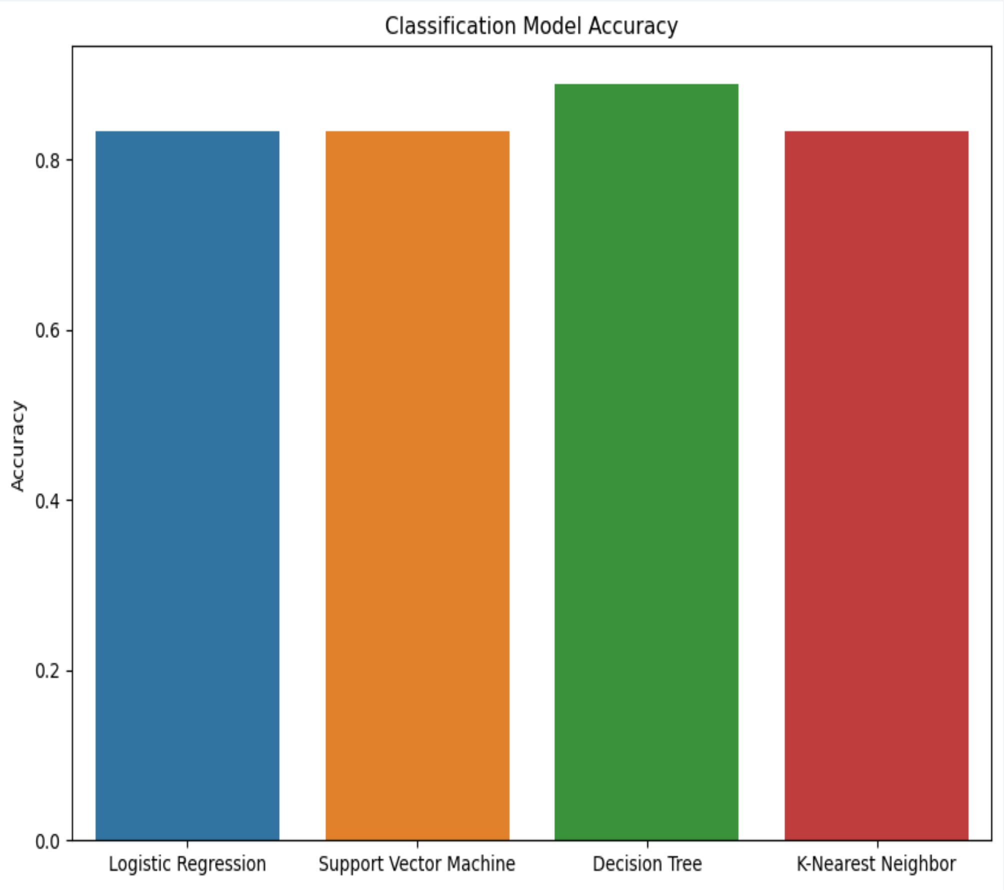
- Payload mass seems to negatively affect launch success especially at higher weights. The higher the mass the likelier the launch outcome is unsuccessful.
  • There seems to be a certain range in payload mass that seems to grant higher success in outcome. See next slide for more ❼

Section 5

Predictive Analysis (Classification)

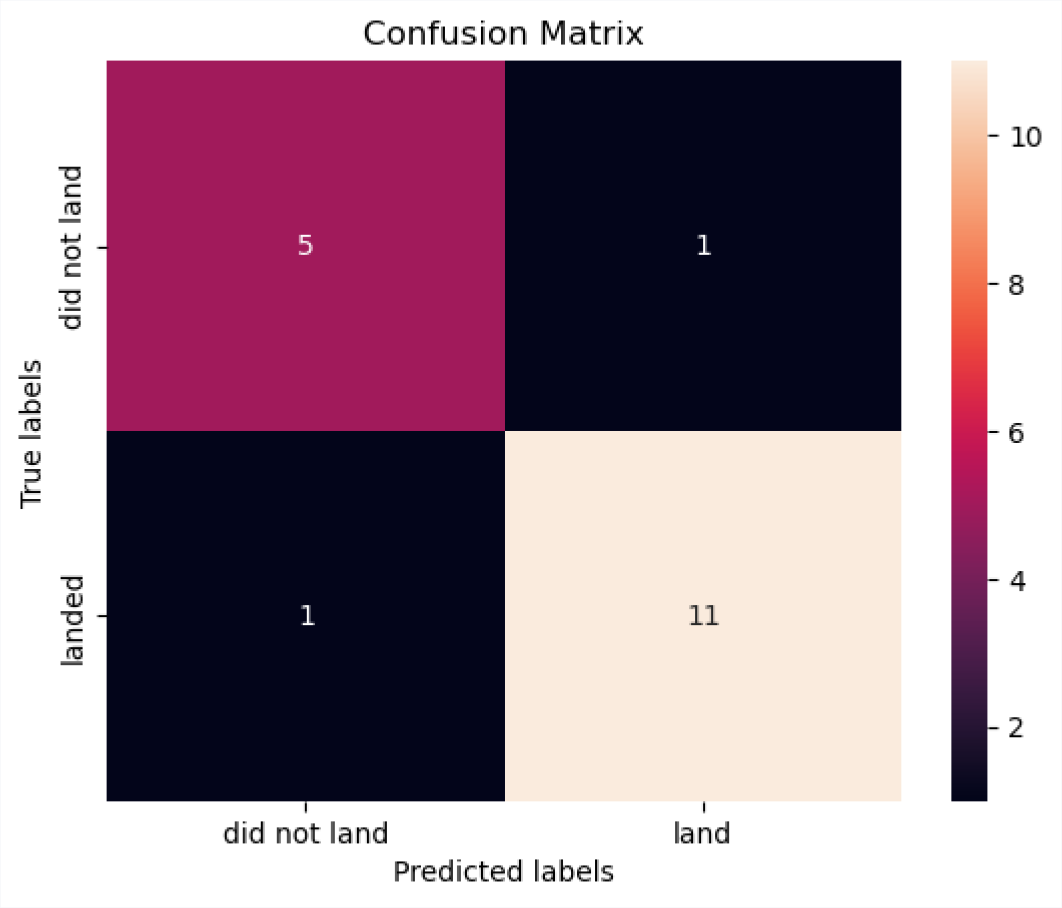
Classification Accuracy

- Classification accuracy on the test data was highest for the decision tree model.
  - Optimal parameters after GridSearchCV:
  - Criterion: Entropy
  - Max Depth: 10
  - Max Features: sqrt
  - Min Samples Leaf: 4
  - Min Samples Split: 2
  - Splitter: Random



Confusion Matrix

- Decision Tree Classifier had an accuracy of 0.89
  - While this classifier might be less optimistic (1 False Negative) it reduced the amount of false positive predictions it made compared to the other models
  - It classified 16 out of 18 outcomes correctly



Conclusions

- There seems to be no relationship between flight number when in GTO orbit
- In payload mass range between 2000 and 5000 kg, booster versions FT (red) and B4 (green) show promising success rates for launch outcome
- When predicting launch outcome a decision tree classifier shows most promising classification accuracy
- As alternative company improve launch outcome success rate in GTO and other orbits where SpaceX is not as developed
- Booster versions FT and B4 show promising results in a certain payload mass range. Improving outcome success rate upon even higher mass and/or showing even better reliability in the aforementioned mass range would also be a way to rival SpaceX

Appendix

- Include any relevant assets like Python code snippets, SQL queries, charts, Notebook outputs, or data sets that you may have created during this project

Thank you!