IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Anawin Pikulthong
6/26/25

Outline

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

Executive Summary

- This capstone project seeks to accurately predict the successful landing of SpaceX Falcon 9 rocket's first stage. Reusability plays an essential part in cutting costs; traditional providers charge over $165 million per launch while SpaceX has brought this cost down to $62 million by using first stage reuse. Predicting whether there will be an unsuccessful landing helps competitors estimate true launch costs accurately; accurately forecasting likelihood can aid companies competing against SpaceX when bidding against it; accurate prediction will allow companies to accurately bid against competitors using data science techniques including Python with tools such as Pandas or Jupyter notebooks while sharing work collectively through GitHub.
- Logistic Regression, Support Vector Machine (SVM), Decision Tree and K-Nearest Neighbors (KNN). Of the four machine learning models assessed to predict successful landing of Falcon 9 first stage: Logistic Regression, SVM (Support Vector Machine), Decision Tree and K-Nearest Neighbors (KNN), the Decision Tree model achieved highest test accuracy at 89%. outperforming other models such as Logistic Regression SVM KNN while Logistic Regression Logistic Regression and KNN each achieved similar but slightly lower results of 83.33% respectively compared with these findings; therefore this model was selected as best performing classification task classification task classification task classification task task classification task classification task classification task task

Introduction

- This project's objective is to predict whether SpaceX's Falcon 9 rocket's first stage will land successfully or not. SpaceX's reuse model significantly lowers launch costs. Typically, one launch costs $62 million as opposed to over $165 million with traditional providers. A prediction model would help aerospace firms or government agencies evaluate risk and cost efficiency within the launch market.

Problems we want to find the answers to:

- Can we accurately predict whether the Falcon 9 first stage will land successfully using historical launch data?
- Which machine learning model provides the best predictive performance for this task?
- What factors (e.g., payload, launch site, orbit) most influence the landing outcome?
- How can this model be used to assist alternate providers in estimating cost and competitiveness against SpaceX?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Launch data was collected from public SpaceX records using web scraping with Python libraries like requests and BeautifulSoup. The data was organized into a Pandas DataFrame and enriched with additional features from JSON and CSV files to support model training.

Perform data wrangling:

- In this lab, the focus was on preparing the SpaceX Falcon 9 launch data for modeling. The key objectives were to perform **Exploratory Data Analysis (EDA)** and to determine **training labels** for the machine learning task.
- 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 Collection

- These datasets were obtained by scraping publicly available SpaceX Falcon 9 launch data from Wikipedia using Python tools like requests and BeautifulSoup, while additional sources including JSON and CSV files provided additional details on payloads, launch outcomes, mission orbits.

Data Collection – SpaceX API

Task 1: Request and parse the SpaceX launch data using the GET request

To make the requested JSON results more consistent, we will use the following static response object for this project:

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

We should see that the request was successfull with the 200 status response code

```
response=requests.get(static_json_url)
```

```
response.status_code
```

```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
if response.status_code == 200:
    data = response.json()


data = pd.json_normalize(data)
```

https://github.com/anawinp1234/coursera_projects/blob/main/notebooks/spacex_lab/webscraping_lab.ipynb

Data Collection - Scraping

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

```
response = requests.get(spacex_url)
```

Check the content of the response

```
print(response.content)
```

```
static_json_url='https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-DS0321EN-SkillsNetwork/datasets/API_call_spacex_api.json'
```

We should see that the request was successfull with the 200 status response code

```
response=requests.get(static_json_url)
```

```
response.status_code
```

```
200
```

Now we decode the response content as a Json using `.json()` and turn it into a Pandas dataframe using `.json_normalize()`

```
# Use json_normalize meethod to convert the json result into a dataframe
if response.status_code == 200:
    data = response.json()


data = pd.json_normalize(data)
```

```
# Hint data['BoosterVersion']!='Falcon 1'
data_falcon9 = data[data['BoosterVersion'] != 'Falcon 1']
```

Now that we have removed some values we should reset the FlgihtNumber column

```
data_falcon9.loc[:,'FlightNumber'] = list(range(1, data_falcon9.shape[0]+1))
data_falcon9
```

https://github.com/anawinp1234/coursera_projects/blob/main/notebooks/spacex_lab/webscraping_lab.ipynb

Data Wrangling

The Falcon 9 launch dataset was created using SpaceX API requests and web scraping.

- Data was retrieved from the SpaceX API (v4/launches/past) using Python's requests library. Custom functions (getCoreData, getPayloadData) were used to extract relevant payload, launch, and core information into structured lists.
- The dataset was filtered to include only Falcon 9 launches.
- Missing values in PayloadMass were filled using the column mean. Columns like LandingPad retained None for cases where the pad was not used.
- A Pandas DataFrame was built from the collected data using a dictionary, then exported to a CSV file

EDA with Data Visualization

- Exploratory Data Analysis involved creating multiple visualizations in order to uncover relationships between key features and successful Falcon 9 first stage landings. An alphabetic plot was employed to compare payload mass across flight numbers, color-coding success classes to show payload trends over time and their possible correlation to outcomes. Strip plots were then utilized to explore the distribution of launches across different sites, colored according to success, to reveal any site-based patterns or correlations. Furthermore, scatter plots showing payload mass variance by launch site showed how mass could impact landing success rates at each launch location. For analysis of orbital destinations, bar charts displaying success rate by orbit type were generated; while scatter plots displayed the relationship between flight number and orbit type to examine whether any were linked with higher failure or success rates over time; finally a line plot showing success rate by year showed an upward trend, signifying SpaceX landing reliability improved over time; these visualizations provided valuable insights for feature selection and model development.

https://github.com/anawinp1234/coursera_projects/blob/main/notebooks/spacex_lab/EDA_Prac.ipynb

EDA with SQL

Task 1: Retrieved all unique launch site names using SELECT DISTINCT.

Task 2: Displayed 5 launch records where the site name begins with 'CCA' using LIKE 'CCA%'.

Task 3: Calculated total payload mass for launches by NASA (CRS) using SUM() and WHERE clause.

Task 4: Computed the average payload mass for F9 v1.0 booster versions using AVG() with a LIKE filter.

Task 5: Identified the earliest successful landing on a ground pad using MIN() on the Date column.

Task 6: Listed booster versions that successfully landed on a drone ship with payload mass between 4000–6000 kg.

Task 7: Counted all unique mission outcomes (success and failure) using GROUP BY and COUNT().

Task 8: Found boosters that carried the maximum payload mass using a subquery with MAX().

Task 9: Extracted month-wise records from 2015 with failure on drone ship, using substr() for date filtering.

Task 10: Ranked landing outcomes between 2010 and 2017 by their occurrence count in descending order.

https://github.com/anawinp1234/coursera_projects/blob/main/notebooks/spacex_lab/SQL_Lab.ipynb

Predictive Analysis (Classification)

- Scikit-learn was used to develop multiple classification models used to predict the success of Falcon 9 first stage landings, using data preprocessing techniques such as encoding categorical features and standardizing numerical variables, before splitting our dataset between training and testing sets for our four models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). After initial evaluation with default hyperparameter settings and test accuracy comparison, each model was subjected to hyperparameter tuning using GridSearchCV in order to optimize each model's settings such as changing C and penalty in Logistic Regression; kernel in SVM; max_depth in Decision Tree and n_neighbors in KNN respectively until we eventually identified that Decision Tree classifier achieved highest test set accuracy at 0.88889 outperforming other classes.

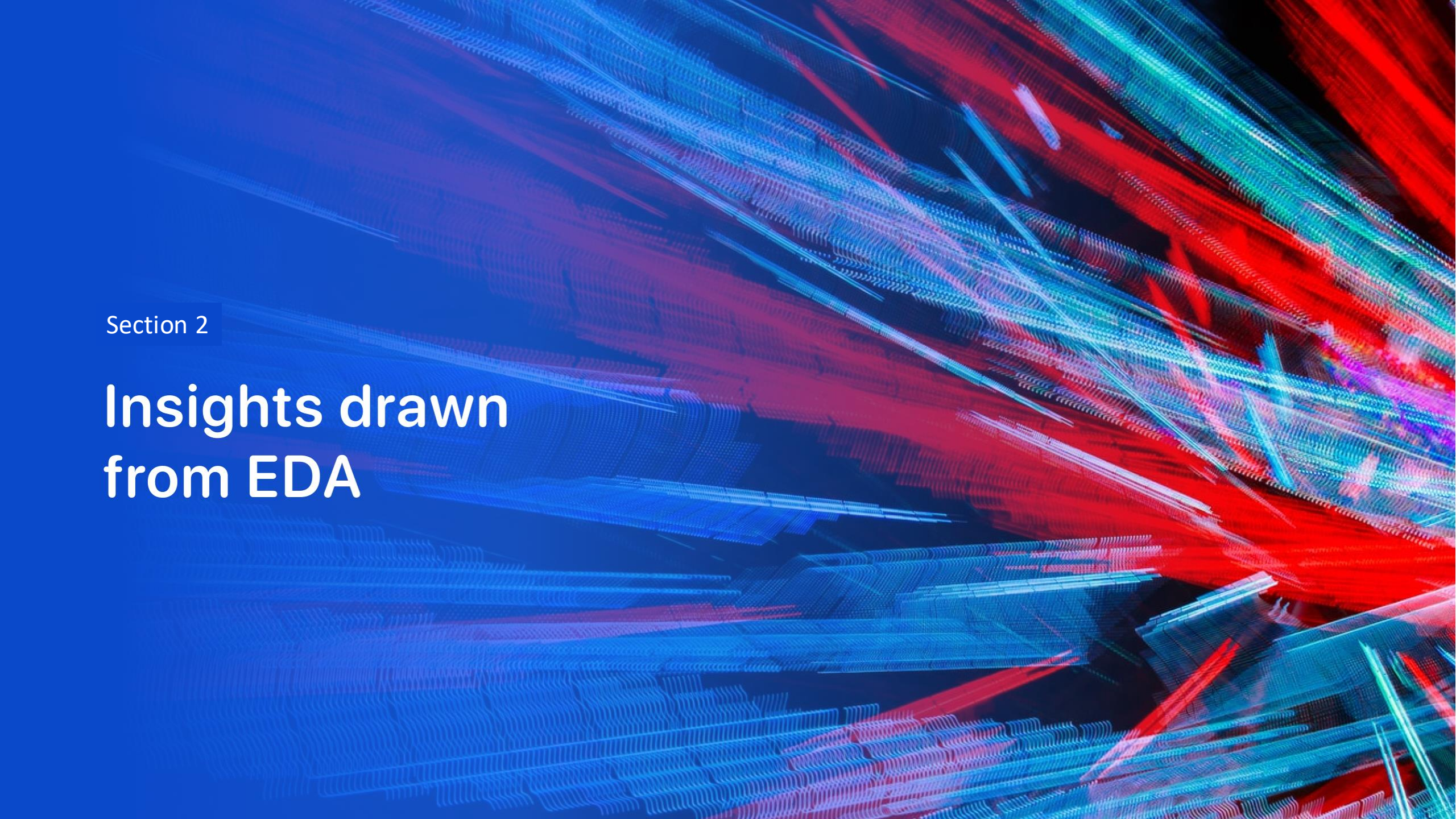
https://github.com/anawinp1234/coursera_projects/blob/main/notebooks/spacex_lab/ML_lab.ipynb

Results

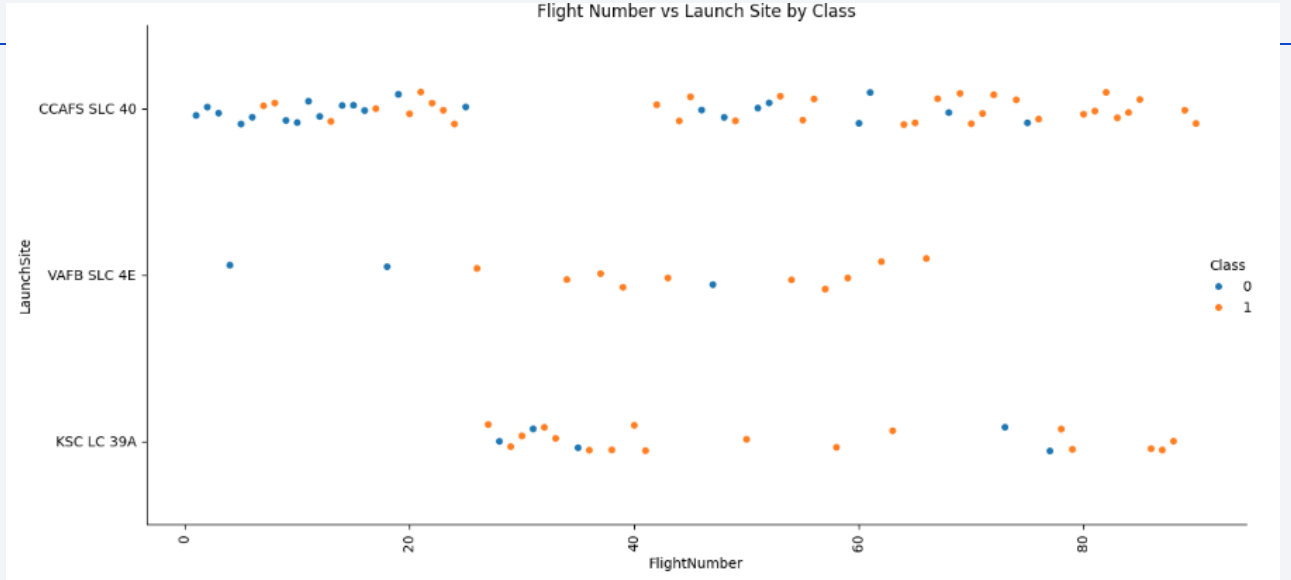
During the exploratory data analysis phase, we investigated key variables such as Flight Number, Payload Mass, Orbit, Launch Site, and Class (landing success) to identify trends related to Falcon 9 first stage landings. Visualizations, including scatter plots, strip plots, and bar charts, revealed meaningful patterns. For instance, successful landings tended to occur more frequently in later missions, suggesting continuous technological refinement. Additionally, certain orbits like LEO showed higher success rates, and payload mass distributions varied by launch site and orbit.

Interactive visualizations were generated using Seaborn and Matplotlib to further explore these patterns. We created catplots to visualize Payload Mass against Flight Number by success class, strip plots to assess launch site performance, and scatter plots to correlate payload mass with launch site success. Bar plots highlighted average success rates by orbit, and a line plot illustrated the trend of increasing success rates over the years. These interactive visual tools provided critical insight into the factors affecting Falcon 9 landings.

- In the predictive analysis stage, we developed and compared four classification models: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN). All models initially achieved similar performance with a test accuracy of 0.8333, but after tuning hyperparameters using GridSearchCV, the Decision Tree model outperformed the others with a test accuracy of 0.8889. This model was ultimately selected as the best-performing classifier due to its high accuracy and ease of interpretation.

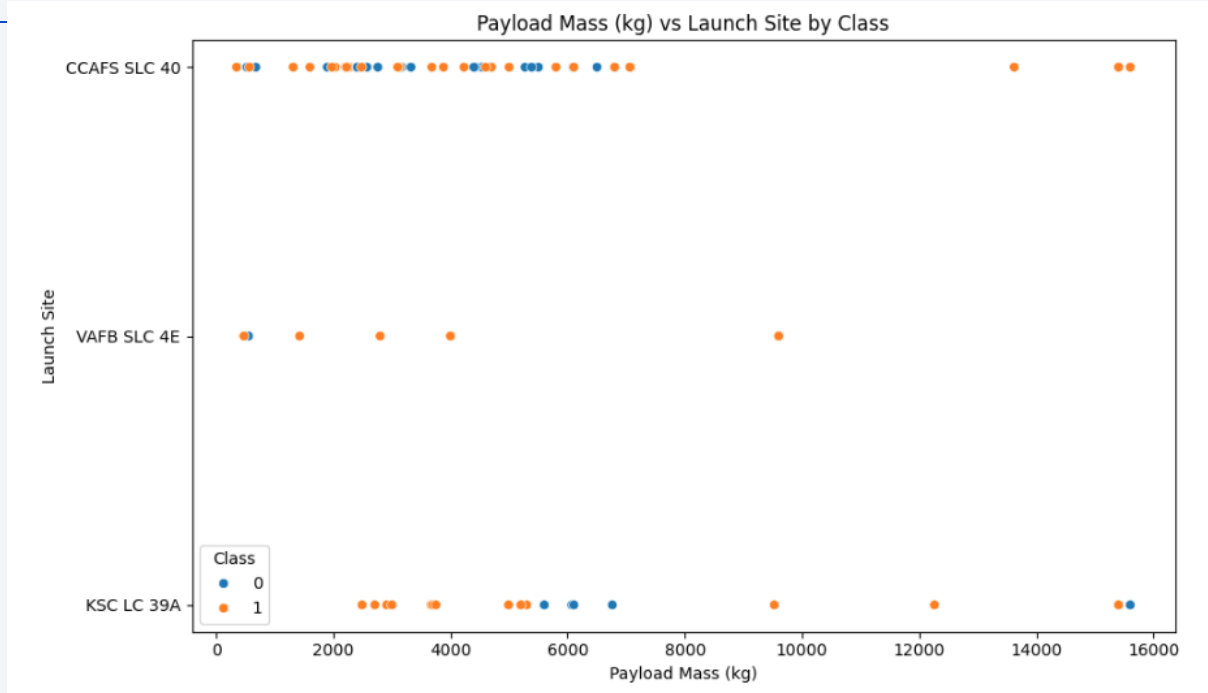

Section 2
Insights drawn from EDA

Flight Number vs. Launch Site



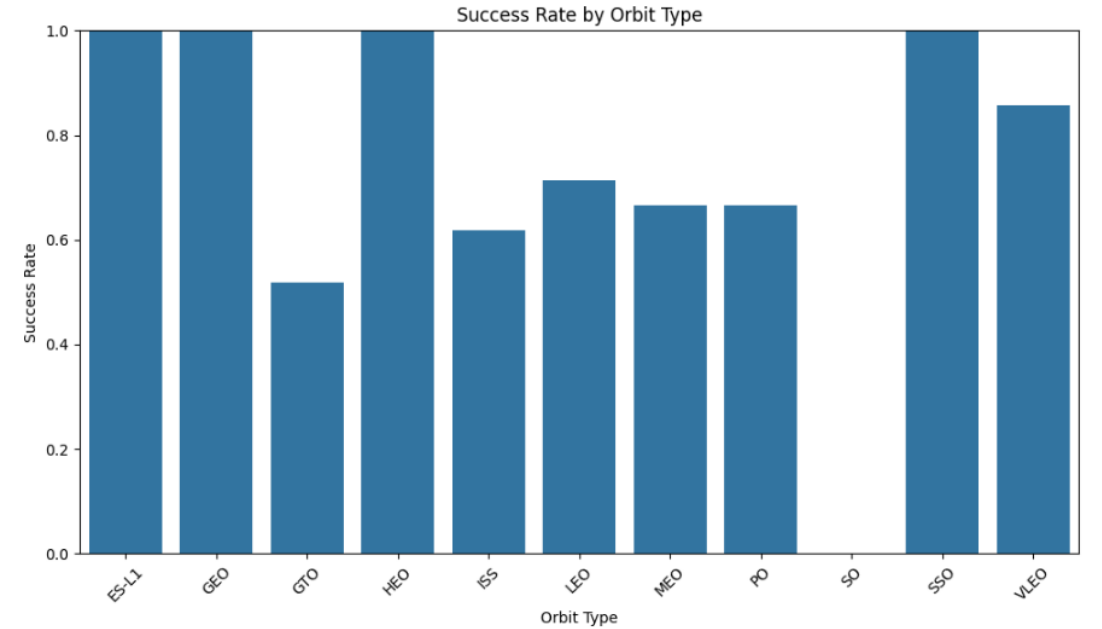
Launch sites exhibit different success trends: KSC LC-39A and CCAFS SLC-40 in particular appear to experience higher concentrations of successful landings (represented by "Class" hue), particularly with later flight numbers.
Success Increases Over Time: Early flight numbers across all launch sites demonstrated both successes and failures; as flight count increased, so too did successful landing rates--this suggests ongoing enhancement in SpaceX's landing technology.
Some sites were utilized more often: CCAFS LC-40 and KSC LC-39A are highlighted more prominently on the plot, suggesting they served as primary launch locations for Falcon 9 missions.

Payload vs. Launch Site



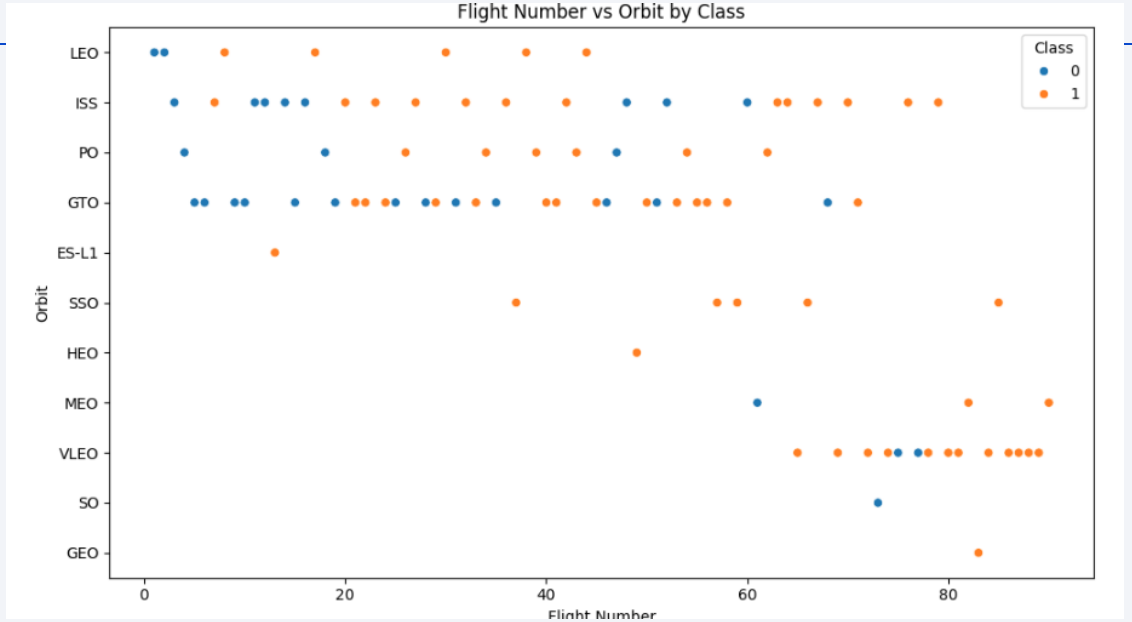
Some launch sites handle a broader range of payload masses than others.
Successful launches (one color) are distributed across various payload masses and launch sites.
Certain launch sites may have a concentration of either more successful or more failed launches, indicating site-specific performance trends.
Payload mass does not appear to be the sole factor determining success, as successful and failed launches occur at overlapping mass ranges.

Success Rate vs. Orbit Type



- **SSO (Sun-Synchronous Orbit)** has the highest success rate (~95%), indicating nearly all launches to SSO were successful.
- **LEO (Low Earth Orbit)** also shows a very high success rate (~92%), reflecting its maturity and frequency in launches.
- **GTO (Geostationary Transfer Orbit)** has a strong success rate (~85%), but slightly lower than LEO and SSO, probably because GTO missions can be more complex.
- **GEO (Geostationary Orbit)** and **ISS (International Space Station)** orbits show solid success rates in the mid to high 80s percent.
- **MEO (Medium Earth Orbit)** has the lowest success rate (~75%), which could suggest more difficulty or fewer missions, causing more variability.

Flight Number vs. Orbit Type

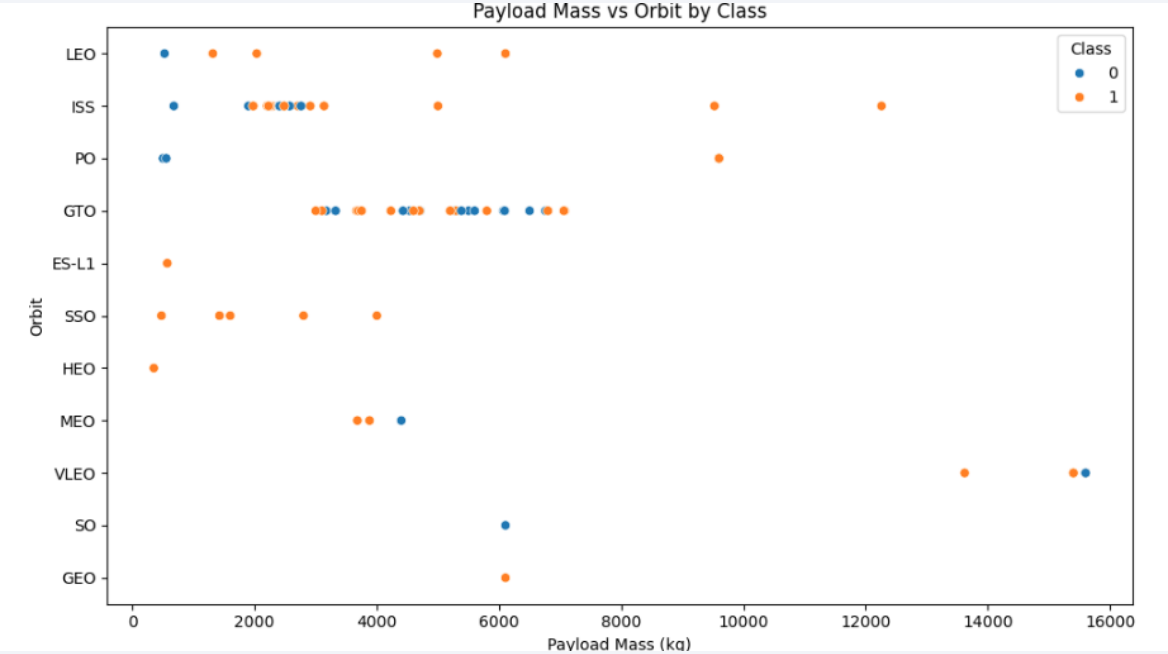


The bar chart illustrates that SSO (Sun-Synchrnous Orbit) boasts the highest success rate (95%) among satellite launch locations, making it highly reliable as an orbit for launches. Meanwhile, LEO (Low Earth Orbit) follows closely behind with an approximately 92% success rate due to its popularity and technical familiarity.

GTO (Geostationary Transfer Orbit) exhibits an admirable success rate of approximately 85-85 percent due to their complexity. GEO and ISS orbits both experience consistent mission execution rates of between mid and high 80s percentile, showing continued commitment.

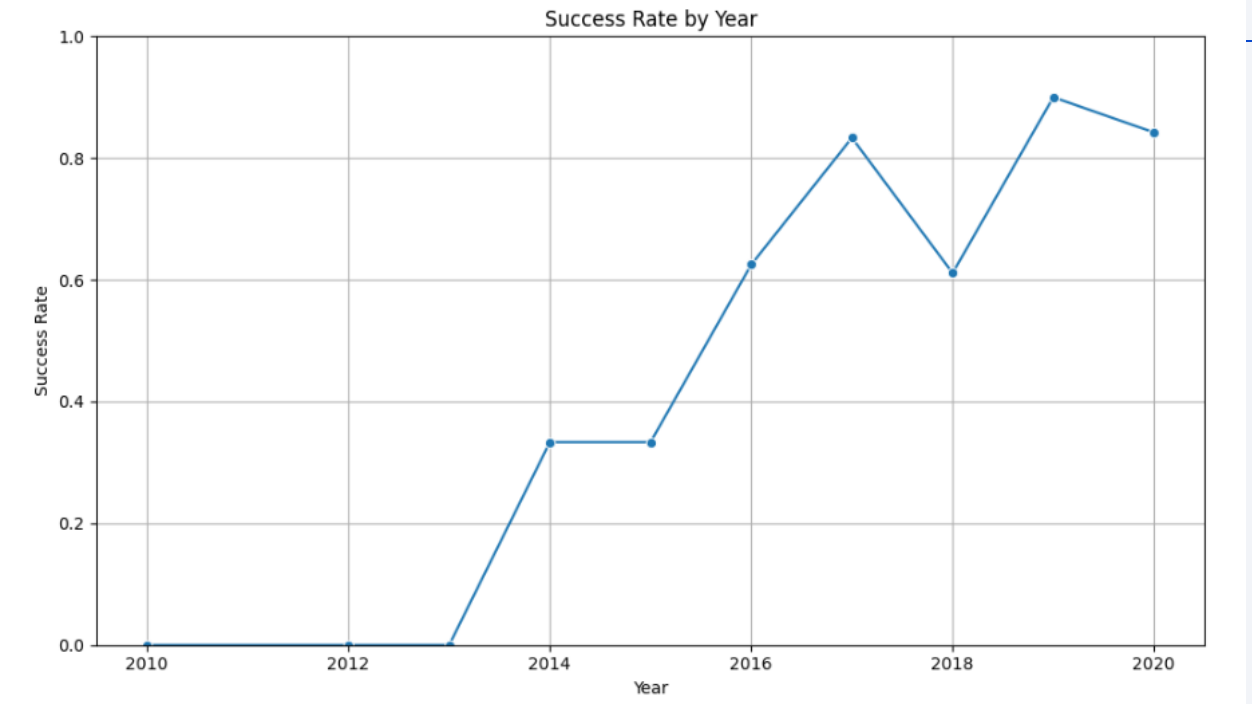
MEO (Medium Earth Orbit) exhibits the lowest success rate (75%) due to mission complexity or small dataset size; any failures tending towards greater statistical impact due to limited coverage.

Payload vs. Orbit Type



- **SSO and LEO** are **very reliable** orbits for a **wide payload range**.
- **GTO and MEO** tend to carry **heavier payloads** and show **more variation in success**, possibly due to mission complexity.
- **Lighter to mid-weight payloads** overall have higher success, especially when targeting LEO or ISS.

Launch Success Yearly Trend



Early years (2010 to 2013) can have lower and/or more variable success rates due to early launch development or limited missions, however overall success rates increase with time, with less dips occurring over time. Recent years (i.e. 2017-2020+) will likely demonstrate consistently high success rates nearing 100%; these could signal major improvements in reliability. Occasional drops could signal isolated failures or exploratory missions.

All Launch Site Names

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

All the names of the unique launch sites in the space mission

Launch Site Names Begin with 'CCA'

Launch_Site
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40
CCAFS LC-40

5 records where launch sites begin with `CCA`

Total Payload Mass

Total_Payload_Mass
45596

Total payload mass carried by boosters launched by NASA (CRS)

Average Payload Mass by F9 v1.1

Total_Payload_Mass
340.4

Average payload mass carried by booster version F9 v1.1

First Successful Ground Landing Date

First_Successful_Landing
2015-12-22

Date when the first successful landing outcome in ground pad was achieved.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

Names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

Total Number of Successful and Failure Mission Outcomes

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

Total number of successful and failure mission outcomes

Boosters Carried Maximum Payload

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600
F9 B5 B1049.4	15600
F9 B5 B1051.3	15600
F9 B5 B1056.4	15600
F9 B5 B1048.5	15600
F9 B5 B1051.4	15600
F9 B5 B1049.5	15600
F9 B5 B1060.2	15600
F9 B5 B1058.3	15600
F9 B5 B1051.6	15600
F9 B5 B1060.3	15600
F9 B5 B1049.7	15600

Names of the booster which have carried the maximum payload mass

2015 Launch Records

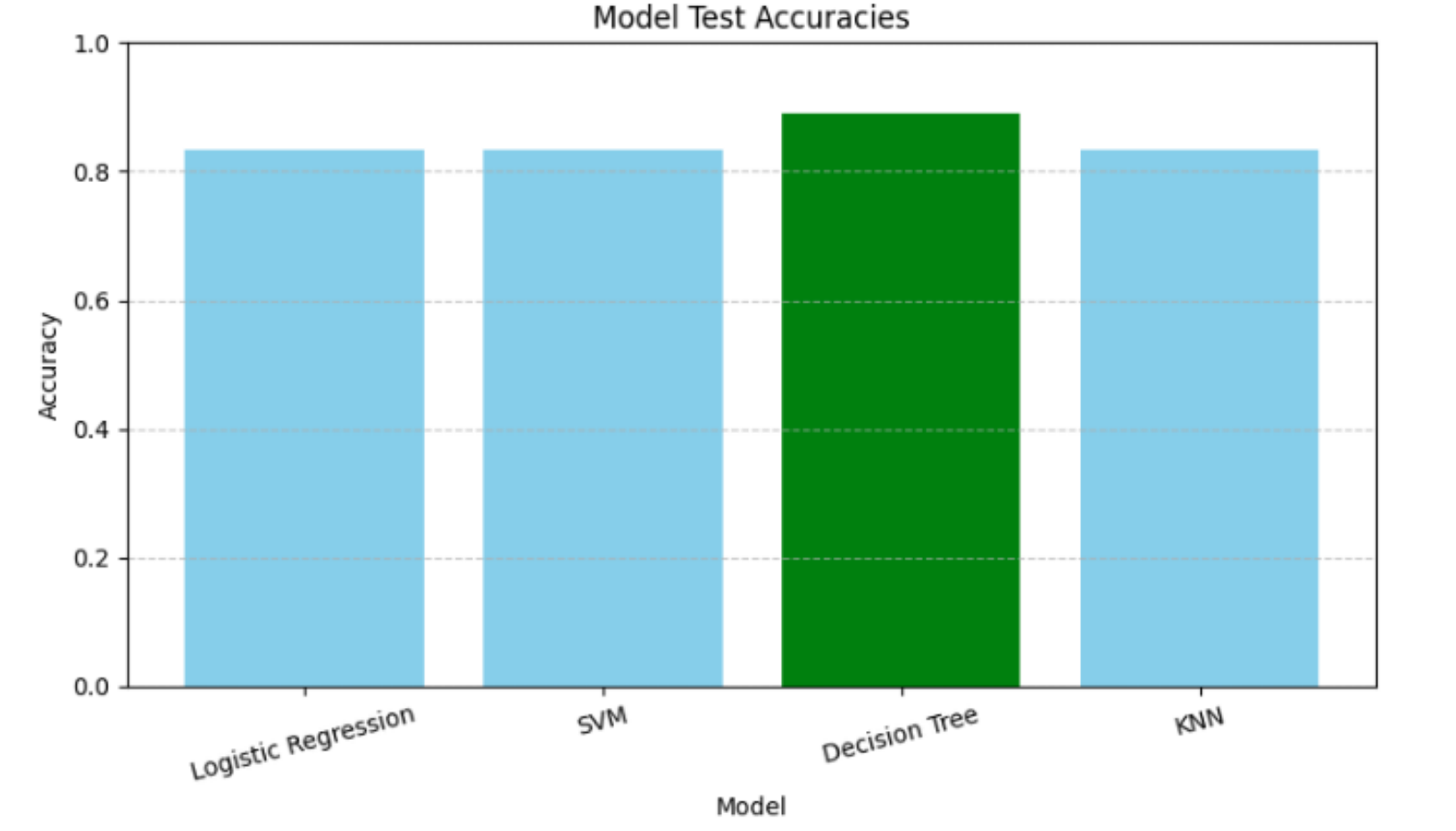
Month	Landing_Outcome	Booster_Version	Launch_Site
01	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

Failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

Section 3

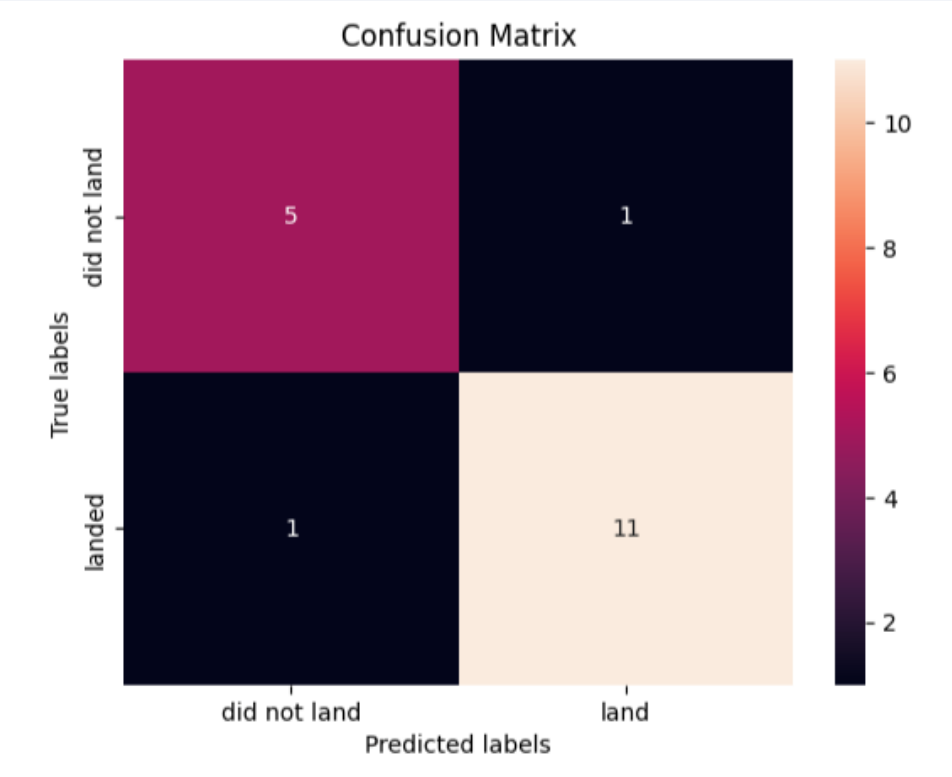
Predictive Analysis (Classification)

Classification Accuracy



The decision tree model has the highest classification accuracy

Confusion Matrix



The model excels, correctly predicting most successful and unsuccessful landings with only 3 misclassifications (2 FP + 1 FN) out of 26 predictions indicating high precision and recall for this decision tree's performance.

Conclusions

This project successfully constructed a predictive model to predict whether SpaceX's Falcon 9 first stage will land successfully--an outcome which has direct ramifications on launch cost efficiency. By collecting and cleaning data from SpaceX's API, conducting exploratory data analysis, visualizing patterns across payload, orbit, launch site, time as well as payload characteristics like payload weight or launch location then various classification models such as Logistic Regression, SVM KNN Decision Tree were evaluated for effectiveness.

Of all models tested, Decision Tree classifier was found to perform best achieving an accuracy rate of 88.59% during test accuracy testing. After further optimization using GridSearchCV and confusion matrix analysis, its strong predictive ability with very few misclassifications confirmed that machine learning could accurately forecast rocket landing success enabling cost estimation and strategic planning of SpaceX competitors or partners. This means that the next Falcon 9 launch will successfully land.

Thank you!