Winning Space Race with Data Science

Lucas Mariétan
19.06.2023

Outline

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

Executive Summary

- Comparison of 4 ML (SVM, logistic regression, KNN, decision tree) algorithms to predict if the first stage of the rocket will land.

- KPI are the same (accuracy 83.3%) for all algorithms -> choose logistic regression because of interpretability and speed.

Introduction

- Private companies are gaining market shares of spaceship launches. Space X has launched rockets with re-usable first stage.
- A company wants to bid against Space X and therefore wants to predict if the first stage of the rocket will land to determine the cost of the launch.

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Collected from Space X API and from wikipedia
- Perform data wrangling
  - Data loaded into pandas dataframes
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Models imported from sklearn library. Data split into test and training sets. Models fitted to training data and prediction done on test data. KPI (accuracy) measured from test data.

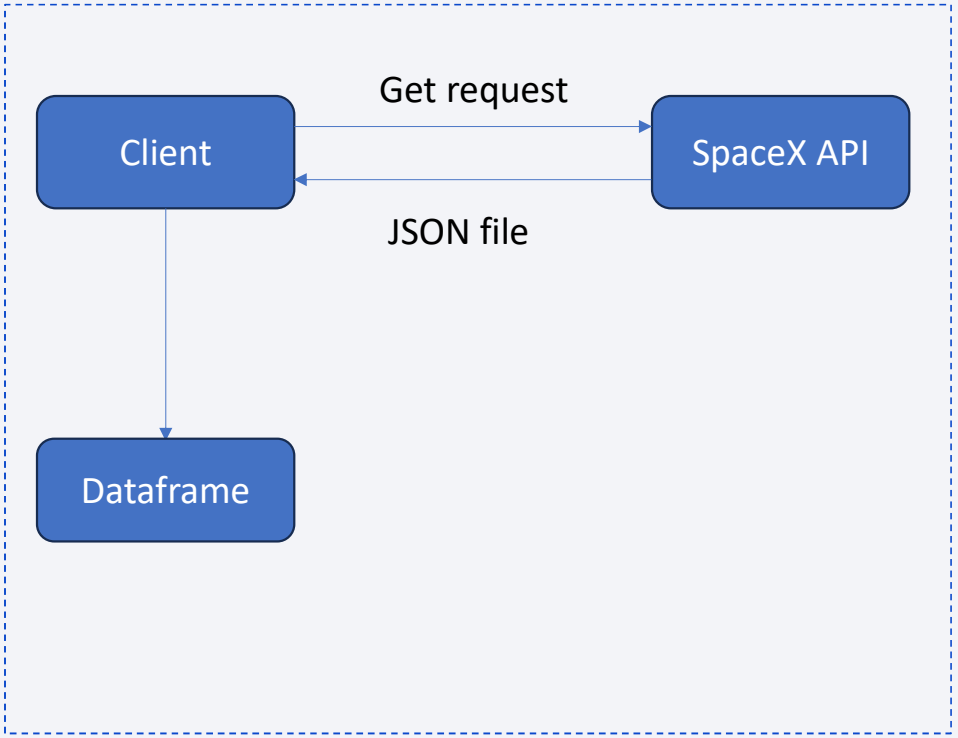
Data Collection

- Data requested from Space X API and scraped from Wikipedia website.

Data Collection – SpaceX API

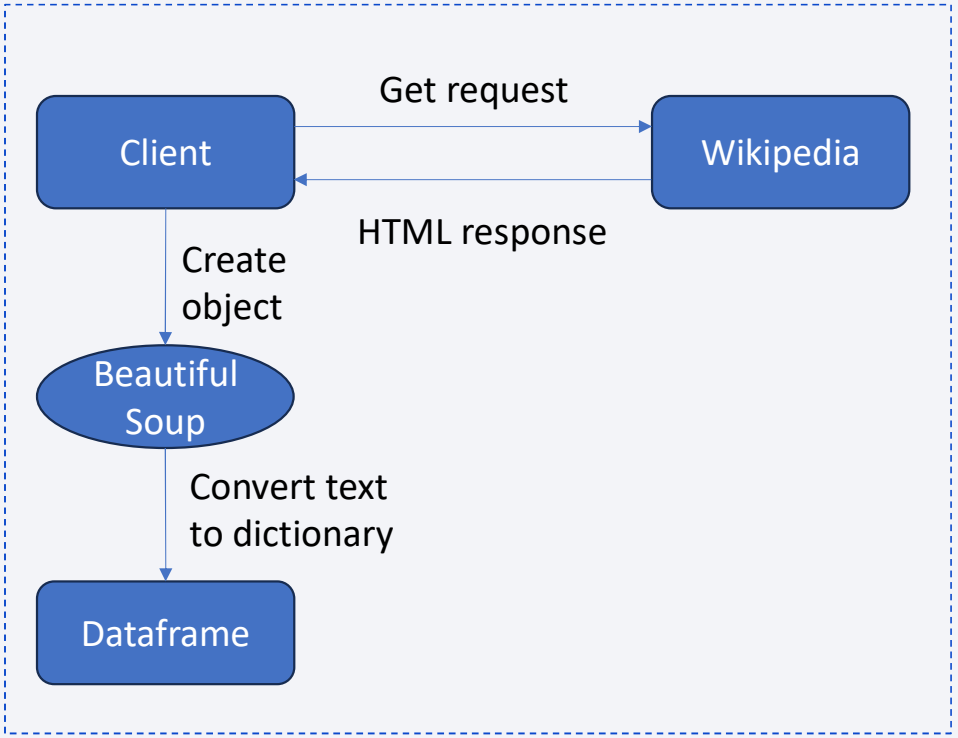
- The client send a get request to the API. The API sends back a JSON file. The JSON is then converted into a dataframe for data analysis.

- Github link: falcon9/jupyter-labs-spacex-data-collection-api.ipynb at main · lmarieta/falcon9 (github.com)



Data Collection - Scraping

- The client send a get request to the wiki page and receive an HTML response. The client parses the HTML with a BeautifulSoup object, extract data in a dictionary and convert to a dataframe.
- Github: falcon9/jupyter-labs-webscraping.ipynb at main · lmarieta/falcon9 (github.com)



Data Wrangling

- Create a binary column representing the outcome of the landing.
- Github: falcon9/data_wrangling.ipynb at main · lmarieta/falcon9 (github.com)

EDA with Data Visualization

- Plot different features against each others to see if they are suitable candidates to build our models (ex Payload Mass vs Filght Number), plot mean success rate as function of target orbit, plot mean success rate along the years. Create binary columns from categorical columns to simplify data analysis.
- Github: falcon9/data-analysis.ipynb at main · lmarieta/falcon9 (github.com)

EDA with SQL

- Select distinct occurences
- Select string containing sub-string
- Select with a where condition
- Select with basic operations (sum, average, …)
- Group by records
- Sub-queries
- Change the order (ascending, descending)
- Github: falcon9/jupyter-labs-data-analysis-sql.ipynb at main · lmarieta/falcon9 (github.com)

Build an Interactive Map with Folium

- Marked all launch sites with circles and markers. Created marker clusters to display failed/successful launches per site. Added mouse position and PolyLine to compute distances.
- Github: falcon9/map_vizualisation.ipynb at main · lmarieta/falcon9 (github.com)

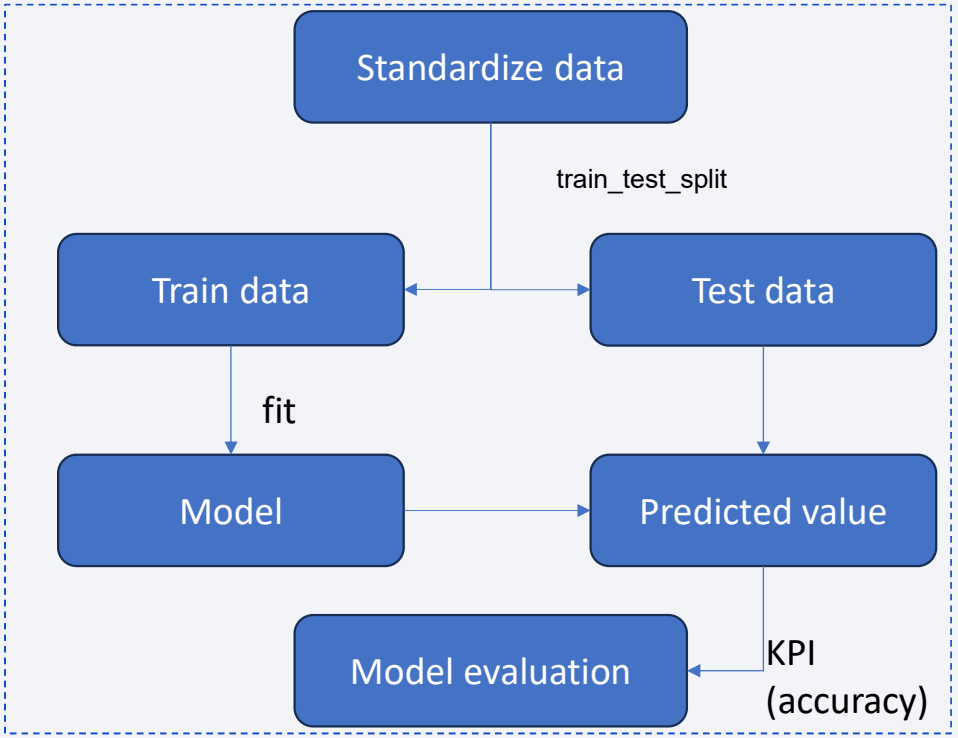
Build a Dashboard with Plotly Dash

- Interactive dashboards to visualize failed/successful launch per sites. Dropdown menu to select the launch site(s). Slider to select the payload range. Scatter chart to show the correlation between payload and launch success.
- Github: falcon9/spacex_dash_app.py at main · lmarieta/falcon9 (github.com)

Predictive Analysis (Classification)

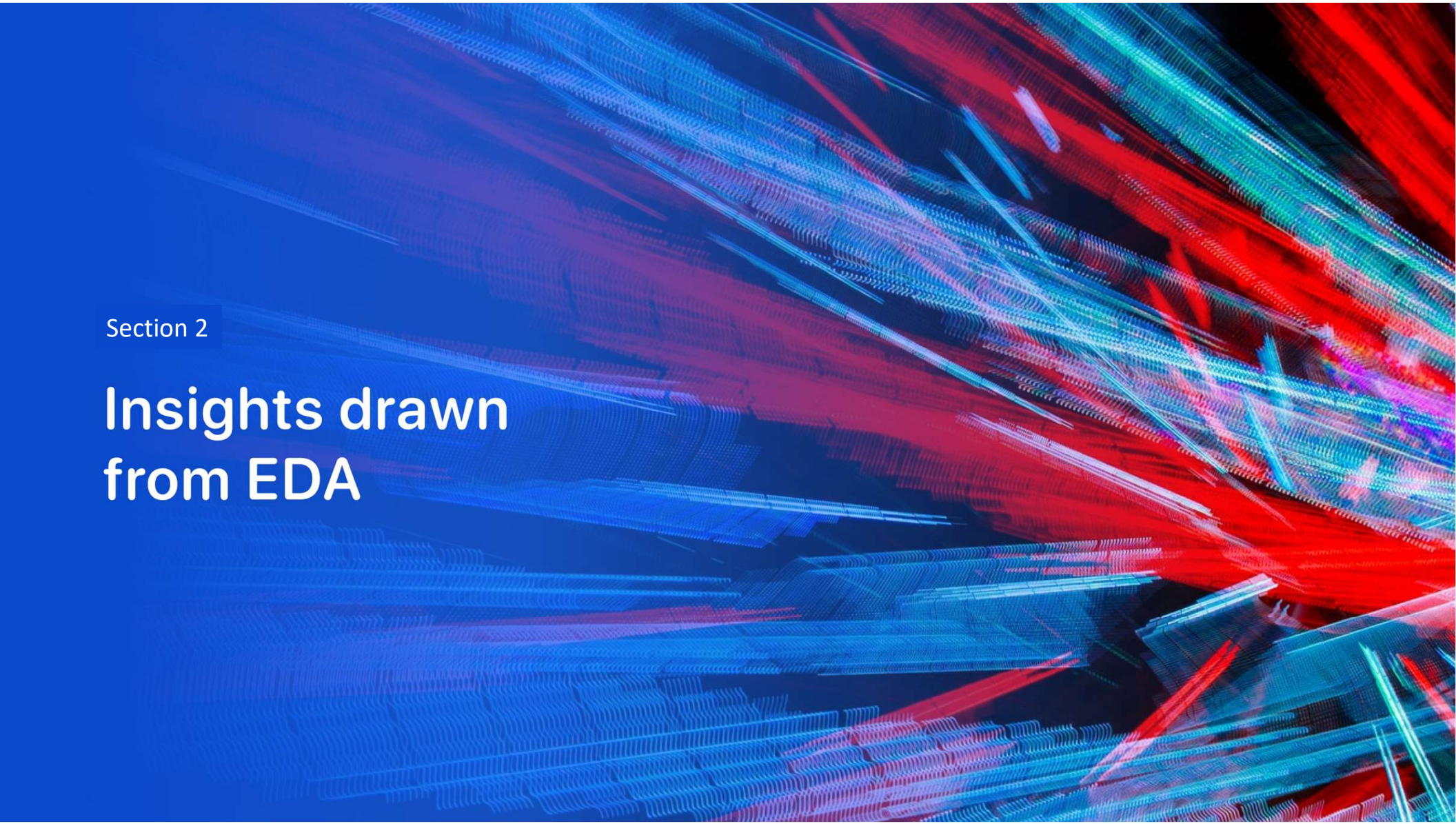
1. Standardize data using StandardScaler
2. Split data into train and tests subsets using train_test_split
3. Fit model to train data using fit and gridsearch to select the hyperparameters.
4. Compute predicted value using the model and test data.
5. Evaluate your model by computing accuracy of predicted values compared to test data. Other KPI should be computed (F1, jaccard, …)

- Github: falcon9/Machine_Learning_Prediction.ipynb at main · lmarieta/falcon9 (github.com)



Results

- EDA: Positive trend for average success rate. Relationships between Payload Mass, Orbit, Launching Site, Flight Number and success rate.
- ML models have all the same performance. They are good to predict the correct outcome in case of successful landing but not in case of failure (i.e. problem of false positives)
- Algorithm can be selected based on the speed or interpretability, i.e we can use the logistic regression.

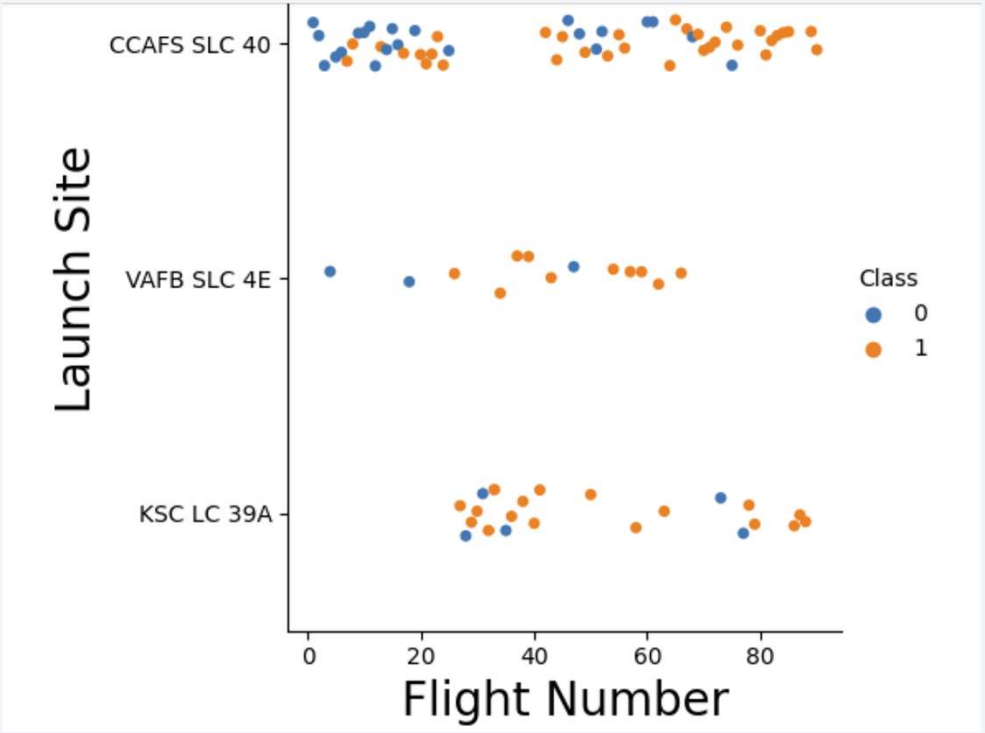


Section 2

Insights drawn from EDA

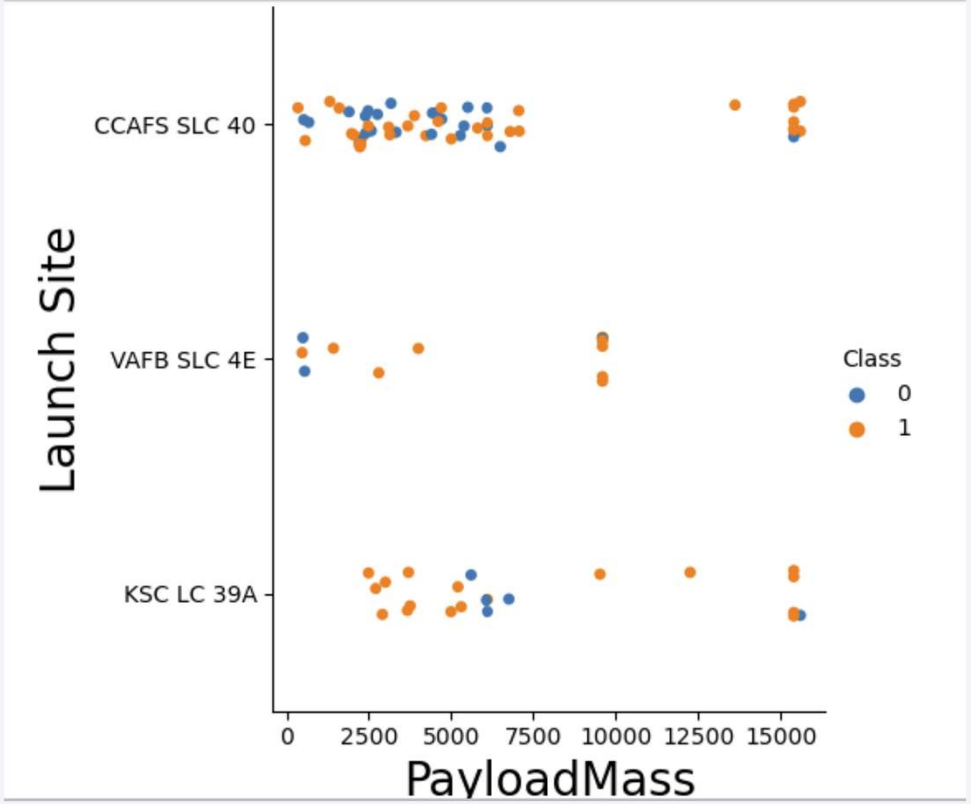
Flight Number vs. Launch Site

- 1 is a successful landing, 0 is failed.

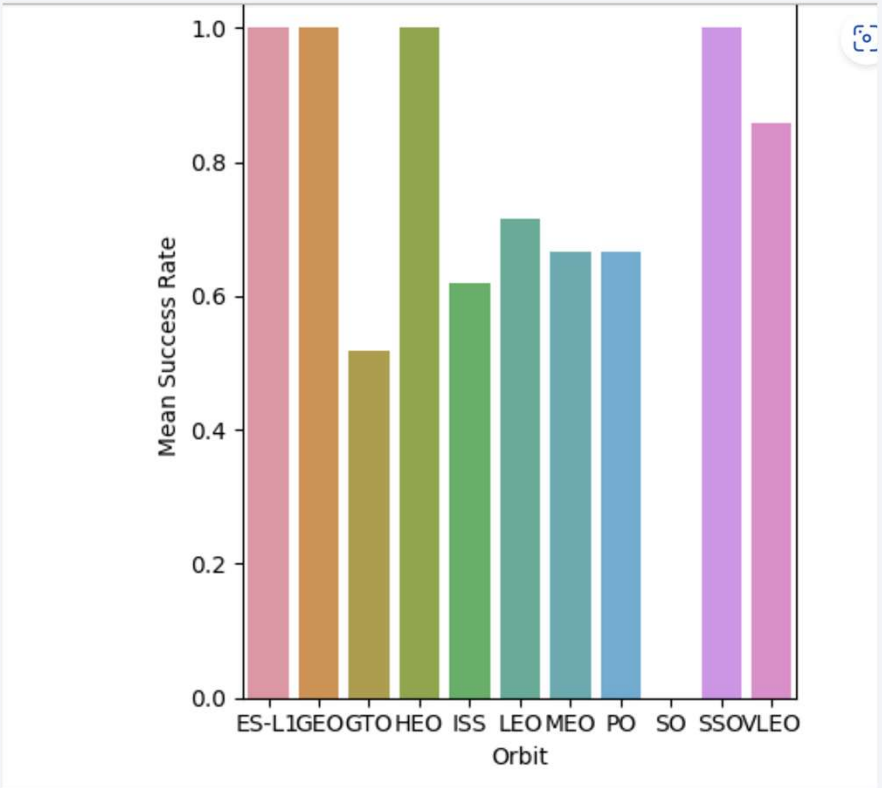


Payload vs. Launch Site

- 1 is a successful landing, 0 is failed.

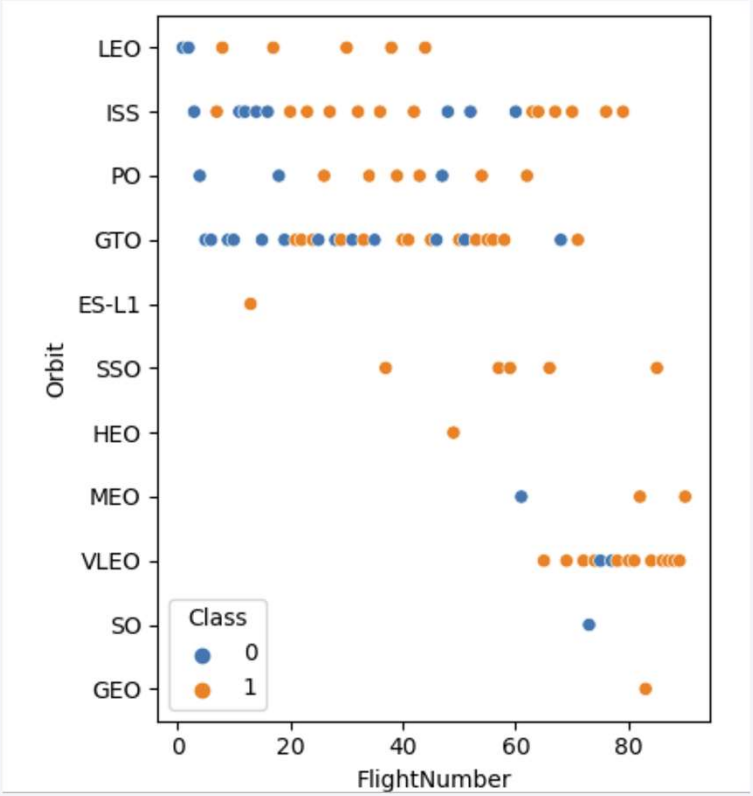


Success Rate vs. Orbit Type



Flight Number vs. Orbit Type

- 1 is a successful landing, 0 is failed.



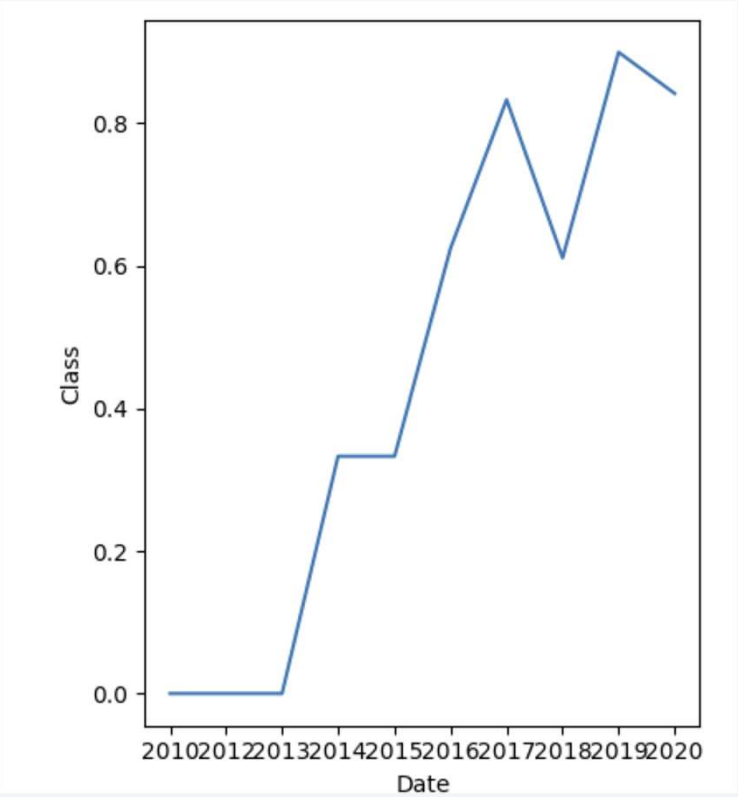
Payload vs. Orbit Type

- Show a scatter point of payload vs. orbit type

- Show the screenshot of the scatter plot with explanations

Launch Success Yearly Trend

- Here class represents the average yearly success rate.



All Launch Site Names

Task 1

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

```
%sql Select DISTINCT("Launch_Site") from "SPACEXTBL";
```

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

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
%sql select * from "SPACEXTBL" where "Launch_Site" like "CCA%" LIMIT 5;
```

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

Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outc
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Su
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Su
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Su
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Su
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Su

Total Payload Mass

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum("PAYLOAD_MASS__KG_") from "SPACEXTBL" where "Customer" = "NASA (CRS)";
```

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

sum("PAYLOAD_MASS__KG_")
45596.0

Average Payload Mass by F9 v1.1

Task 4

Display average payload mass carried by booster version F9 v1.1

```
%sql select "Booster_Version",avg("PAYLOAD_MASS__KG_") from "SPACEXTBL" where "Booster_Version" like "F9 v1
```

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

Booster_Version	avg("PAYLOAD_MASS__KG_")
F9 v1.1	2928.4

First Successful Ground Landing Date

Task 5

List the date when the first succesful landing outcome in ground pad was acheived.

Hint:Use min function

```
%sql select min("DATE") from "SPACEXTBL" where "Landing_Outcome" = "Success (ground pad)";
```

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

min("DATE")
01/08/2018

Successful Drone Ship Landing with Payload between 4000 and 6000

Task 6

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

```
%sql select "Booster_Version" from "SPACEXTBL" where "Landing_Outcome" = "Success (drone ship)" AND "PAYLOA
```

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

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

Total Number of Successful and Failure Mission Outcomes

Task 7

List the total number of successful and failure mission outcomes

```
]: %sql select "Mission_Outcome",COUNT(*) from "SPACEXTBL" group by "Mission_Outcome"
```

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

Mission_Outcome	COUNT(*)
None	898
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
%sql select "Booster_Version","PAYLOAD_MASS__KG_" from "SPACEXTBL" where "PAYLOAD_MASS__KG_" = (SELECT max(
```

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

Booster_Version	PAYLOAD_MASS__KG_
F9 B5 B1048.4	15600.0
F9 B5 B1049.4	15600.0
F9 B5 B1051.3	15600.0
F9 B5 B1056.4	15600.0
F9 B5 B1048.5	15600.0
F9 B5 B1051.4	15600.0
F9 B5 B1049.5	15600.0
F9 B5 B1060.2	15600.0
F9 B5 B1058.3	15600.0
F9 B5 B1051.6	15600.0
F9 B5 B1060.3	15600.0
F9 B5 B1049.7	15600.0

2015 Launch Records

```
%sql select substr("Date", 4, 2) as month,"Landing_Outcome","Booster_Version","Launch_Site" from "SPACEXTBL
```

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

month	Landing_Outcome	Booster_Version	Launch_Site
02	No attempt	F9 v1.1 B1014	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40
04	No attempt	F9 v1.1 B1016	CCAFS LC-40
06	Precluded (drone ship)	F9 v1.1 B1018	CCAFS LC-40
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
11	Controlled (ocean)	F9 v1.1 B1013	CCAFS LC-40
12	Success (ground pad)	F9 FT B1019	CCAFS LC-40

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

Task 10

Rank the count of successful landing_outcomes between the date 04-06-2010 and 20-03-2017 in descending order.

```
%sql SELECT "Landing_Outcome", COUNT(*) as count FROM "SPACEXTBL" WHERE "DATE" >= '04/06/2010' AND "DATE" <
```

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

Landing_Outcome	count
No attempt	9

Section 3

Launch Sites Proximities Analysis

Launch site locations

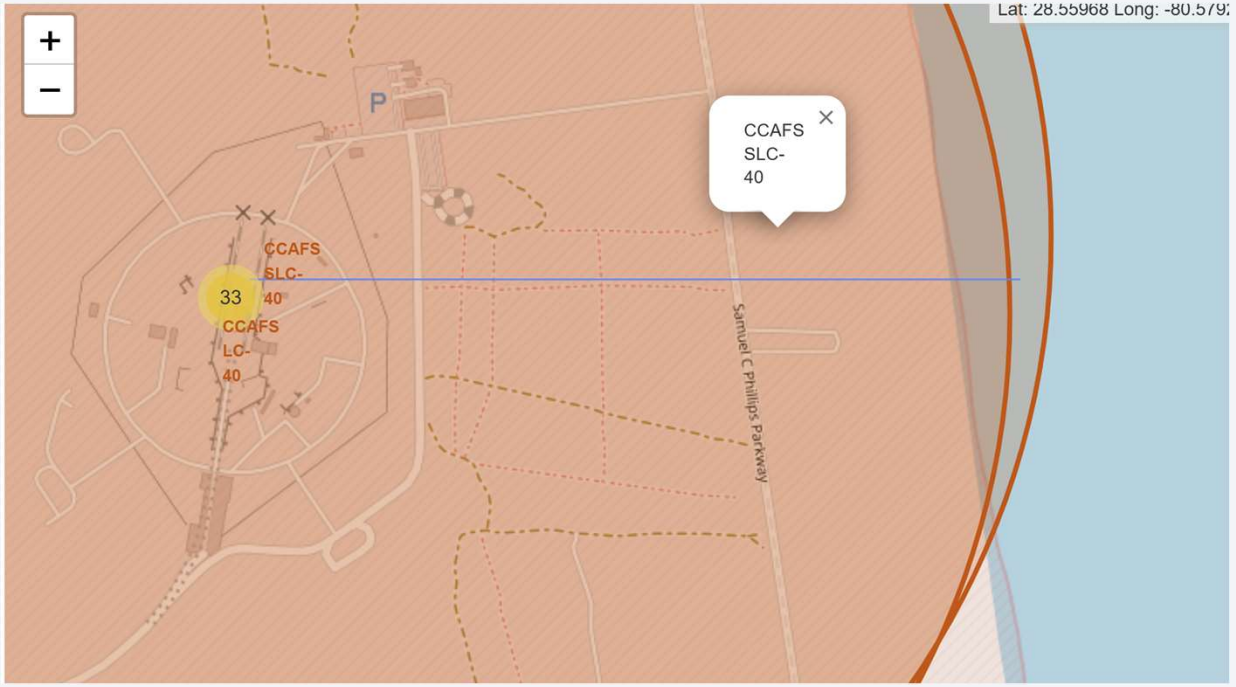
- 3 launch sites are in close vicinity of each other (in Florida) and are superimposed on the map

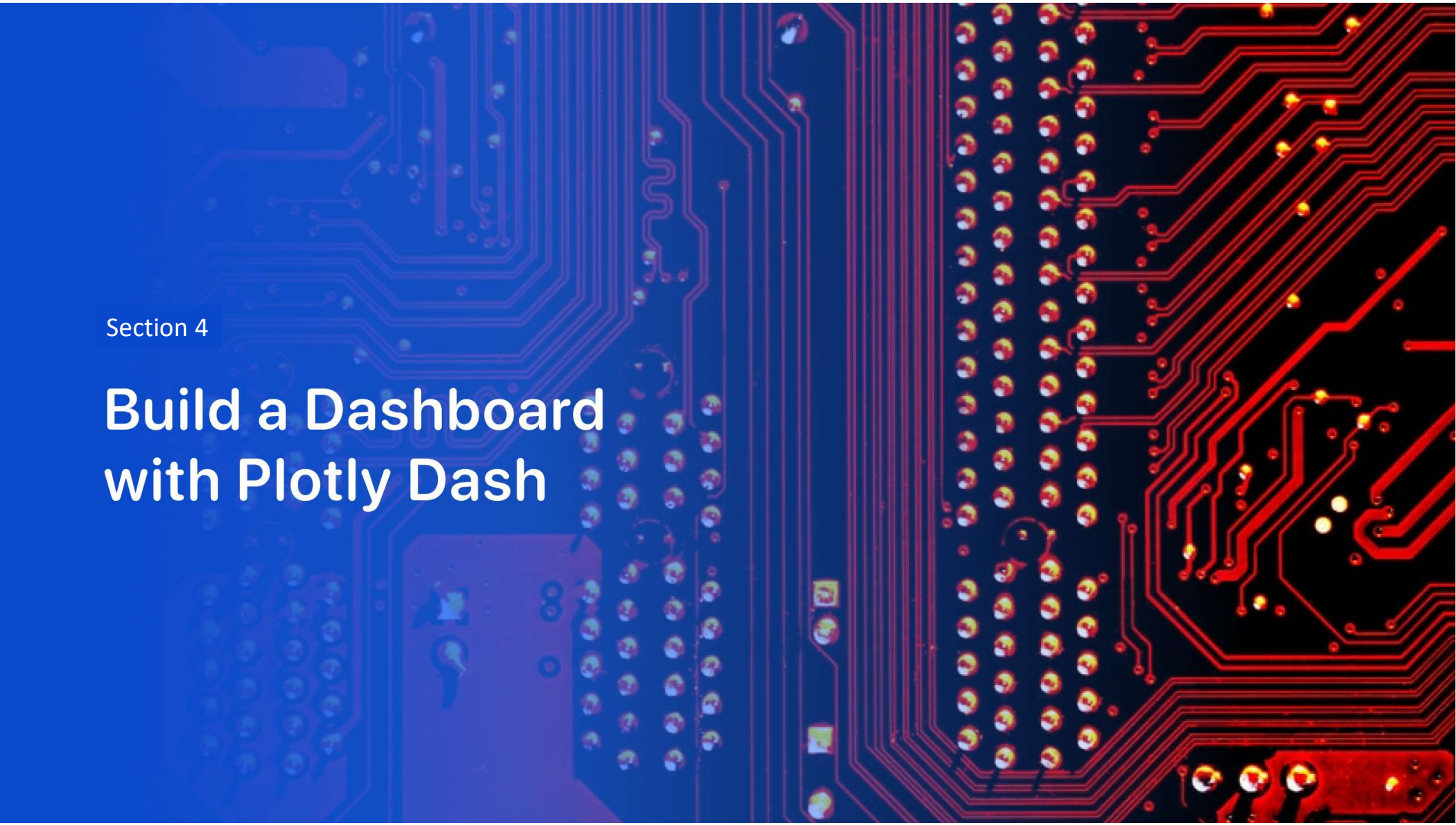
Successfully/failed landings per site

- Here the yellow circles are the number of launch per site.
- The map is interactive and clicking on a yellow circle display the green and red icons, which represent successful landings, failed respectively.

Polyline for distance measurement

- Here we add a blue line to compute a distance between two points





Section 4

Build a Dashboard with Plotly Dash

TBD code works, screenshots to be provided

- Replace <Dashboard screenshot 1> title with an appropriate title
- Show the screenshot of launch success count for all sites, in a piechart
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 2>

- Replace <Dashboard screenshot 2> title with an appropriate title
- Show the screenshot of the piechart for the launch site with highest launch success ratio
- Explain the important elements and findings on the screenshot

<Dashboard Screenshot 3>

- Replace <Dashboard screenshot 3> title with an appropriate title
- Show screenshots of Payload vs. Launch Outcome scatter plot for all sites, with different payload selected in the range slider
- Explain the important elements and findings on the screenshot, such as which payload range or booster version have the largest success rate, etc.

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- No need to use a bar chart as all models have the same accuracy…

TASK 12

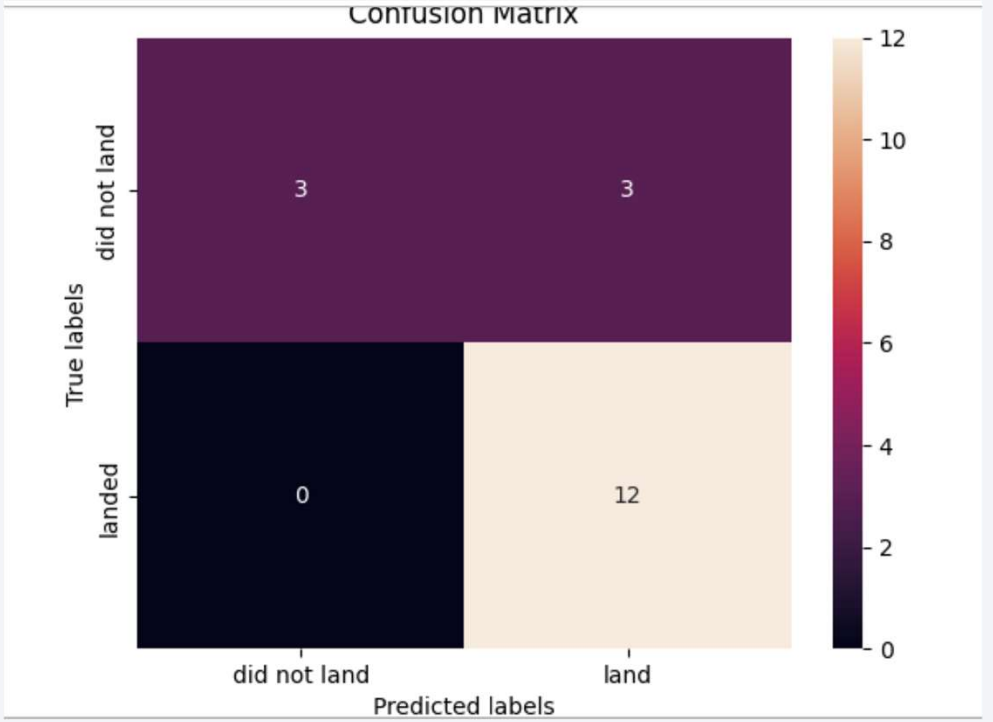
Find the method performs best:

```
print(accuracy_log)
print(accuracy_svm)
print(accuracy_tree)
print(accuracy_knn)
```

```
0.8333333333333334
0.8333333333333334
0.8333333333333334
0.8333333333333334
```

Confusion Matrix

- All models have the same confusion matrix. We notice that the models lack accuracy regarding false positives.



Conclusions

- All models are equivalent. They are accurate to predict the landing outcome in the case where the landing is successful.
- SpaceX is in a positive trend regarding successful landings.

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!