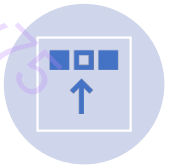
IBM Capstone Project

@Niicky75

Outline



Executive summary

Introduction

Methodology

Results

Discussion

Conclusion

Executive Summary

We Will predict if the SpaceX Falcon 9 first stage will land successfully using machine learning algorithms

The Project wil include:

Data collection, Wrangling, Formatting

Exploratory data análisis

Interactive data visualiation

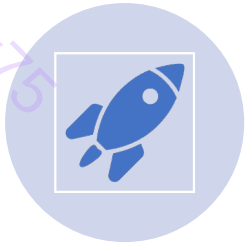
Machine learning prediction

It exists a correlation between rocket launche features and the outcome of the launch (success or failure).

Decision tree is the best machine learning algorithm to predict if the Falcon 9 first stage Will land successfully.

Introduction

In this presentation, we Will predict if the Falcon 9 first stage Will land successfully. SpaceXX advertises Falcon 9 rocket launches for 62 million dollars instead of 165 million dollars if case of not reusable rockets. Therefore we would be able to determine if the the first stage Will land and determine the cost of the launch. This information can be used by competitors sucha s Space Y.



We will analyse the impact of given features and data as the payload mass, the orbit type or the launch site… on the rocket landing.

Methodology

Data collection , wrangling, formatting using:

SpaceX API

Web Scrapping with Beautiful Soup

Exploratory Data Analysis using:

Pandas and Numpy

SQL

Data Visualization using:

Matplotlib and Seaborn

Folium for geographical map

Plotly Dash

Machine Learning predictions using:

Logistic regression

Support Vector Machine (SVM)

Decision tree

K-nearest neighbors (KNN)

Methodology : Data Collection, Wrangling, Formatting

The API used is https://api.spacexdata.com/v4/rockets/

The API provides data about rocket launches by SpacX, the data is therefore filtered to include only Falcon 9 Launches.

Every missing value in rows is replaced by the mean of each columna, to avoid altering values during statistic calculations.

We end up with 90 rows as instances and17 columns as features.

	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005	-80.577366	28.561857
6	3	2013-03-01	Falcon 9	677.0	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007	-80.577366	28.561857
7	4	2013-09-29	Falcon 9	500.0	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003	-120.610829	34.632093
8	5	2013-12-03	Falcon 9	3170.0	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004	-80.577366	28.561857

Methodology : Data Collection, Wrangling, Formatting

The data is web scrapped from https://en.Wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

The data only contains data about Falcon 9 launches.

We end up with 121 rows as instances and 11 columns as features.

	Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
0	1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success\n	F9 v1.0B0003.1	Failure	4 June 2010	18:45
1	2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
2	3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt\n	22 May 2012	07:44
3	4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success\n	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
4	5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success\n	F9 v1.0B0007.1	No attempt\n	1 March 2013	15:10

Methodology : Data Collection, Wrangling, Formatting

- The data is later processed to avoid missing entries
- Categorical feature values are one-hot encoded
- An additional column ‘Class’ is added as 0 if the launch failed and 1 if the launch succeeded
- We end up with 90 rows and 83 columns

Methodology: Exploratory Data Analysis

Function from Pandas and Numpy are used used to extract information such as:

Number of Launches

Number of occurrence of each orbit

Number of occurrence of each mission outcome

The data are queried using SQL to answer the following information:

Names of the unique launch sites in the space mission

Total Payload Mass carried by boosters launched by NASA (CRS)

The average payload mass carried by boosters version F9 v1.1

Methodology : Data Visualization

Matplotlib and Seaborn libraries are used to visualize the data through Scatterplot, Bar charts and Line charts to analyse the relationship between:

Flight number and launch site

Payload mass and launch site

Orbit type and success launch rate

Folium library is used to visualize map through interactive maps. It is used to:

Mark launch sites on a map

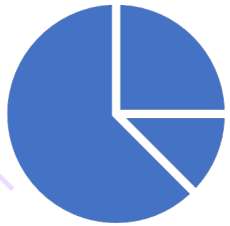
Mark the succeeded and failed launches of each site on a map

Mark the distance between a launch site and its proximities such as nearest city, railway or highway

Folium

Methodology : Data Visualization

Plotly Dash is used to display an interactive site with input toggle availability using a dropdown menu and range slider.



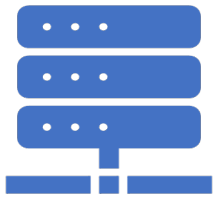
Using a pie chart and a scatter plot, the site shows:

The total success rate for each site

The correlation between payload mass and mission outcome for each site

Methodology : Machine Learning Predictions

Scikit-learn library is used to creat machine learning models.



The following steps are followed:

Data standardizations.

Data splitting between train set and test set.

Creation of machine learning models as:

- Logistic Regression
- Support Vector Machine (SVM)
- Decision Tree
- K-Nearest Neighbors (KNN)

Fit the model in the training set.

Find the best combination of hyperparameters for each model.

Evaluate the models based on their accuracy score and confusion matrix.

Results

The results are split in 5 sections:

- SQL (EDA using SQL)
- Matplotlib and Seaborn (EDA with visualization)
- Folium
- Plotly Dash
- Predictive analysis

Class = 1 == Successful launch

Class = 0 == Failed Launch

Results: SQL (EDA with SQL)

Names of unique launch sites in the space mission

5 records where launch beggin with ‘CCA’

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

DATE	time_utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
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	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Results: SQL (EDA with SQL)

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

Total payload mass by NASA (CRS)
45596

Average payload mass carried by booster version F9 v1.1

Average payload mass by Booster Version F9 v1.1
2928

Date of first successful landing outcome in ground pad was achieved

Date of first successful landing outcome in ground pad
2015-12-22

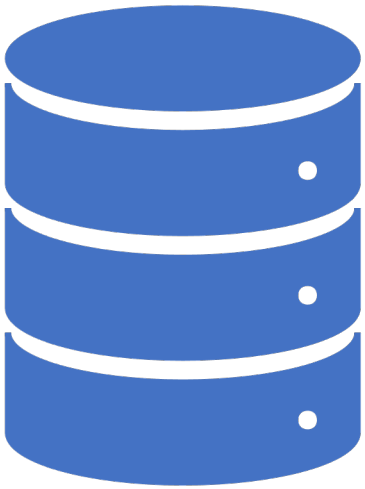
Results: SQL (EDA with SQL)

Name of boosters wich have success in drone ship and have payload mas 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 success and failure outcomes

number_of_success_outcomes	number_of_failure_outcomes
100	1



booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

Results: SQL (EDA with SQL)

Names of boosters versions that have carried the maximum payload mass

Results: SQL (EDA with SQL)

- Failed landing outcomes in drone ship, their booster versions, and launch site

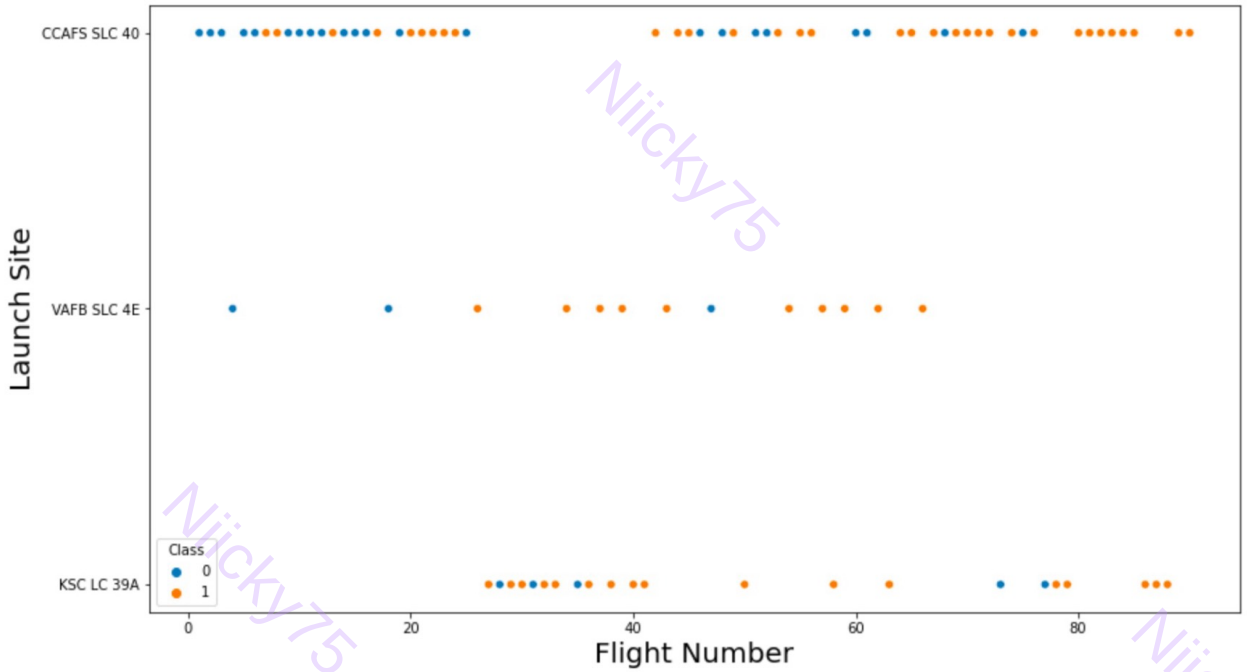
DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

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

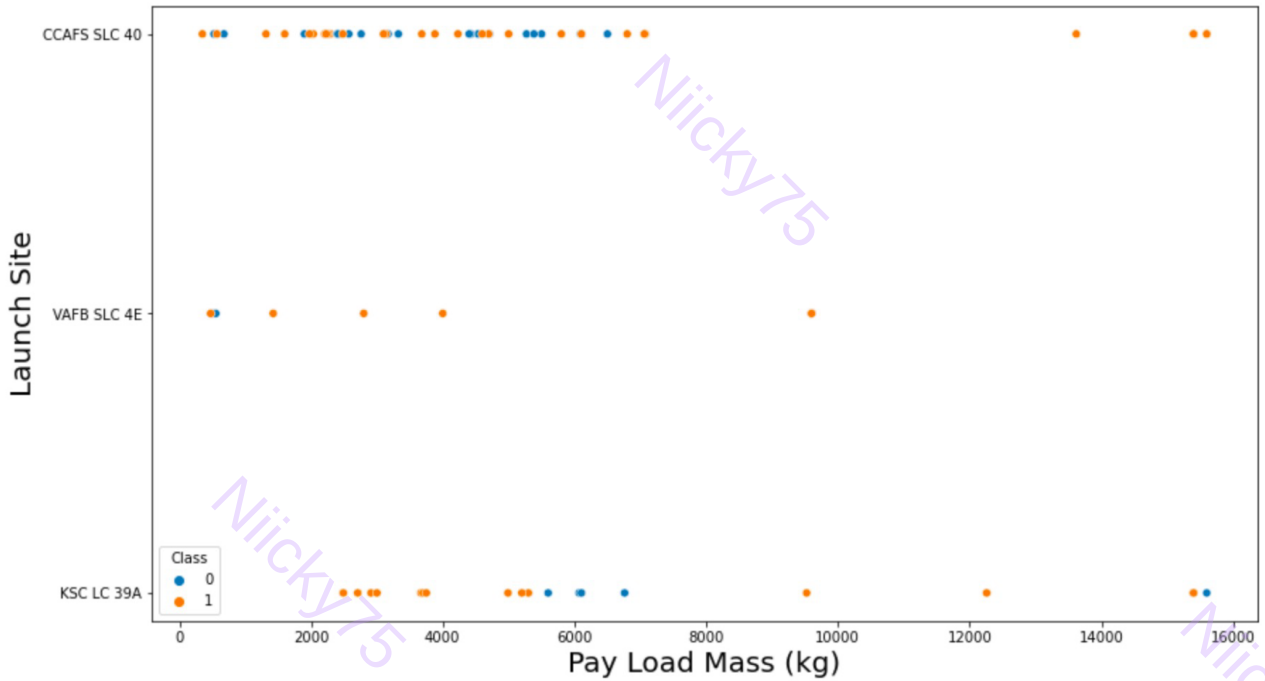
Results: Matplotlib & Seaborn (EDA with Visualization

- Relationship between flight number and launch site



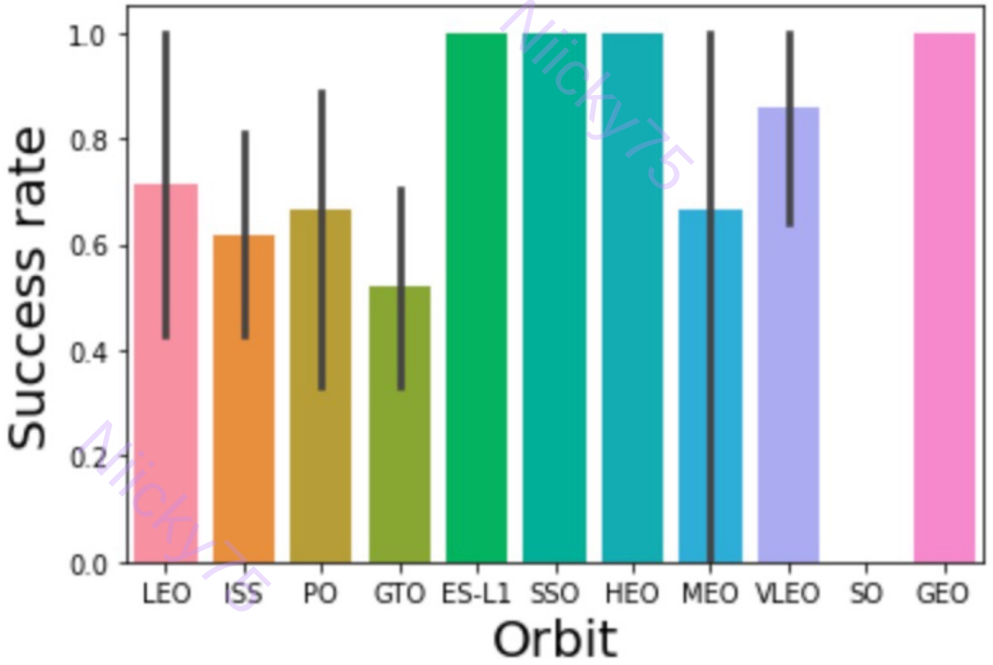
Results: Matplotlib & Seaborn (EDA with Visualization

- Relationship between payload mass and launch site



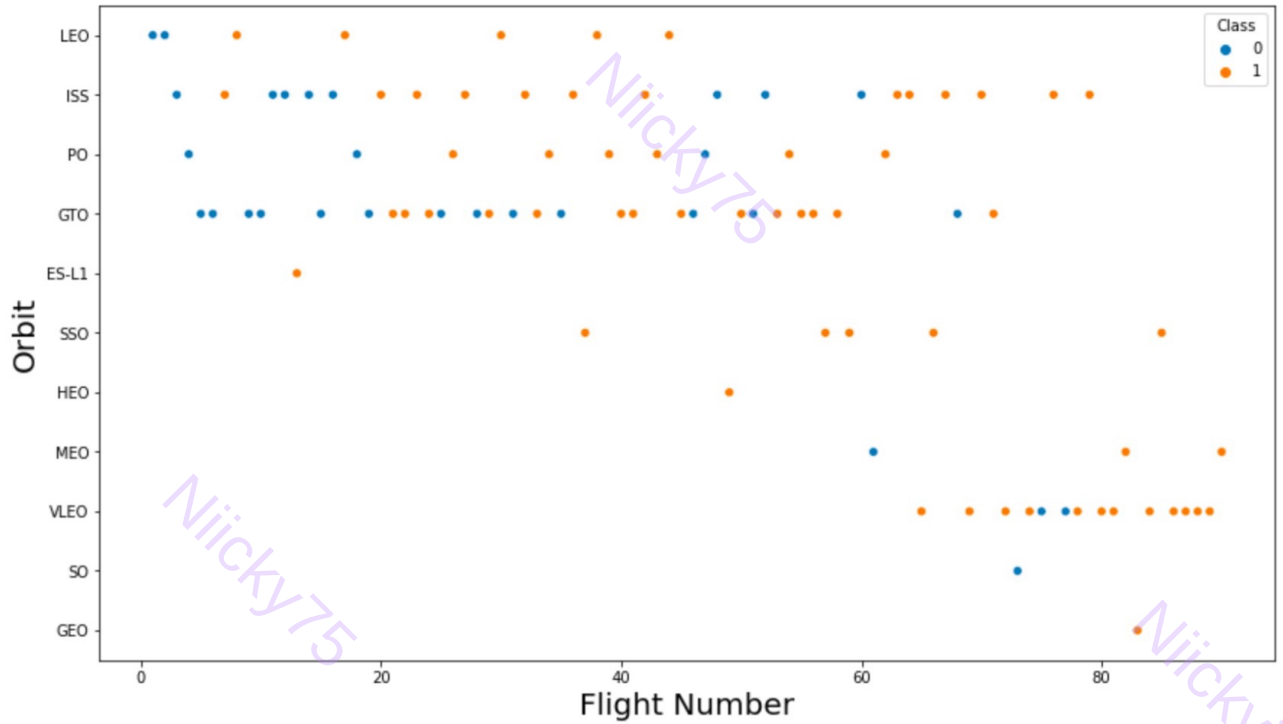
Results: Matplotlib & Seaborn (EDA with Visualization

- Relationship between success rate and orbit type



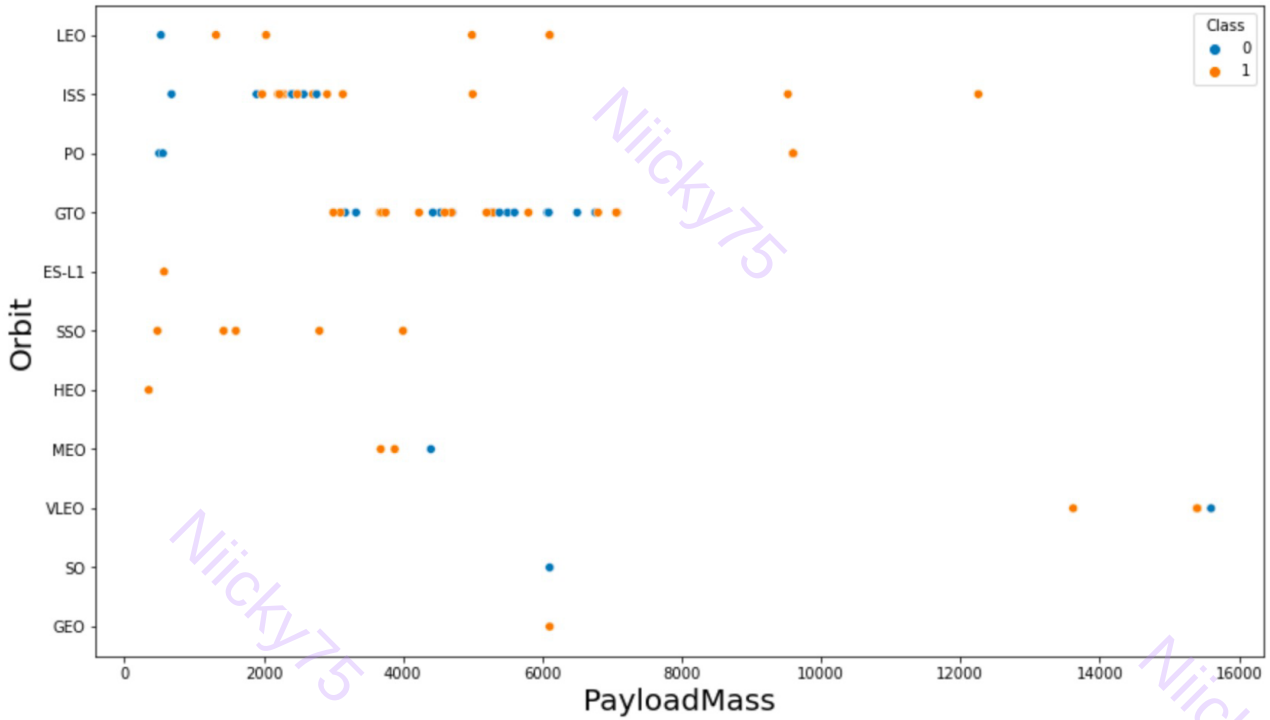
Results: Matplotlib & Seaborn (EDA with Visualization

- Relationship between flight number and orbit type



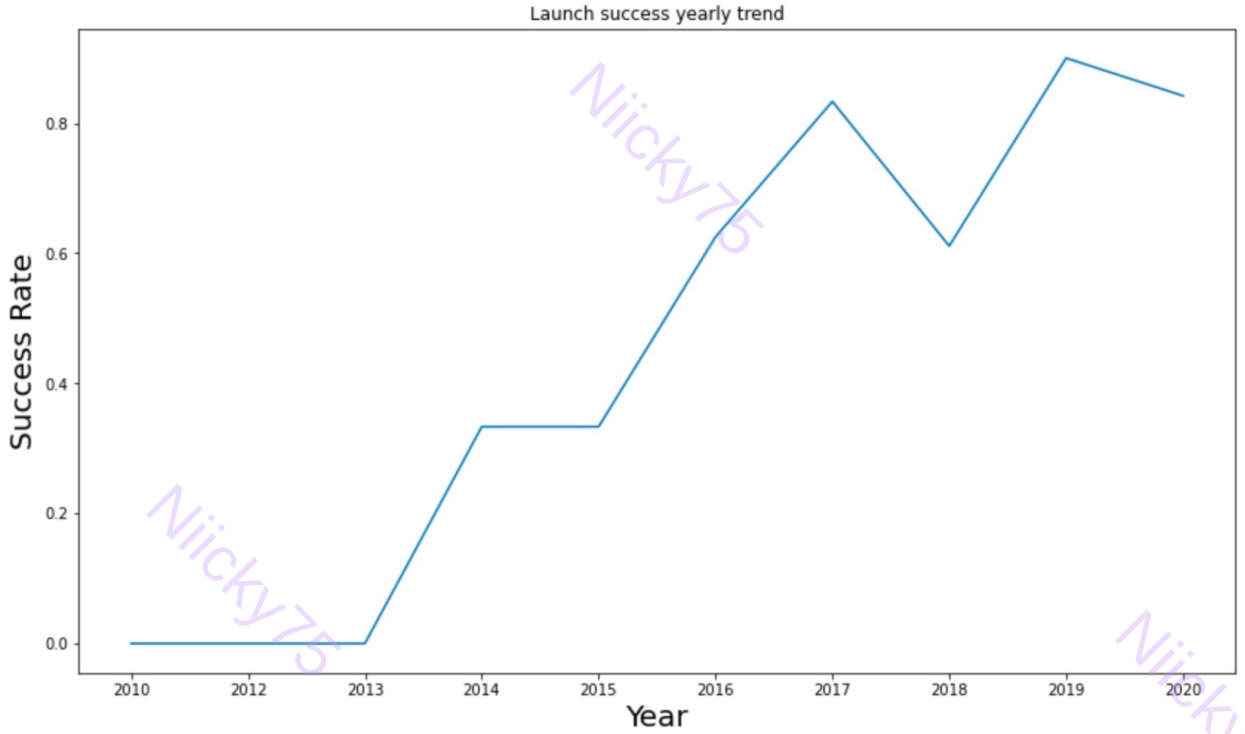
Results: Matplotlib & Seaborn (EDA with Visualization

- Relationship between payload mass and orbit type



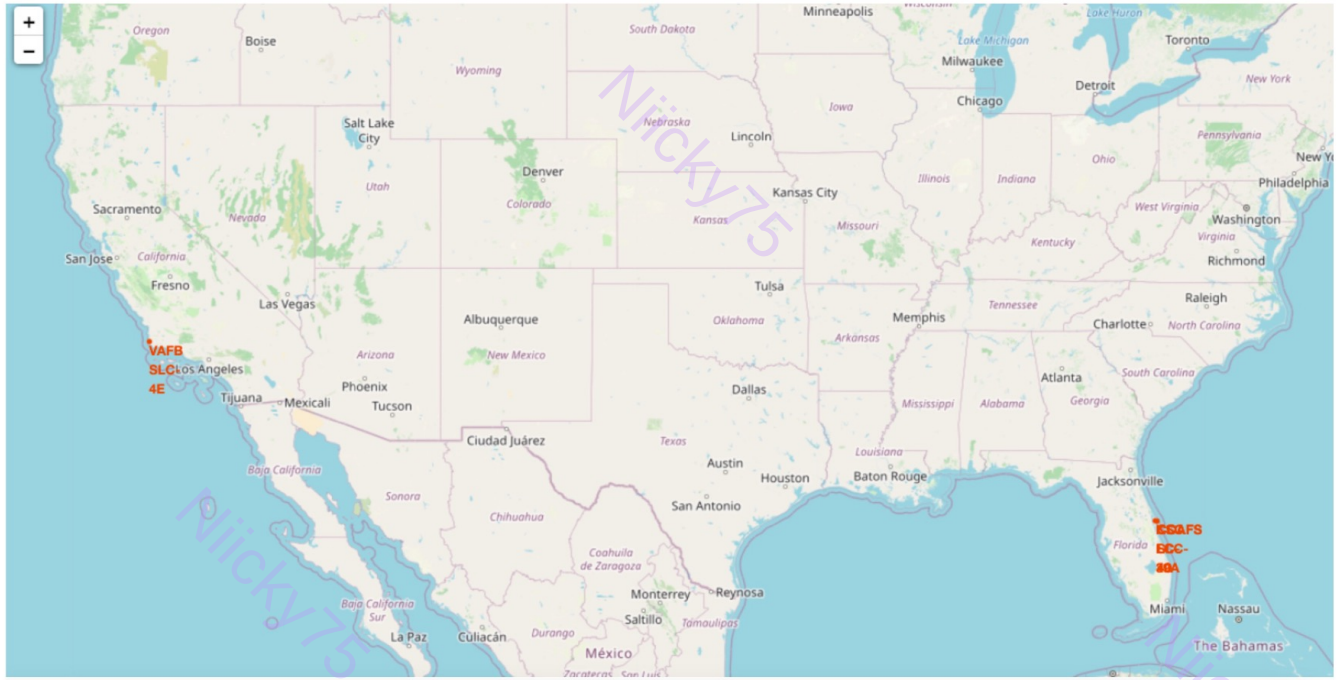
Results: Matplotlib & Seaborn (EDA with Visualization

- Launch success yearly trend



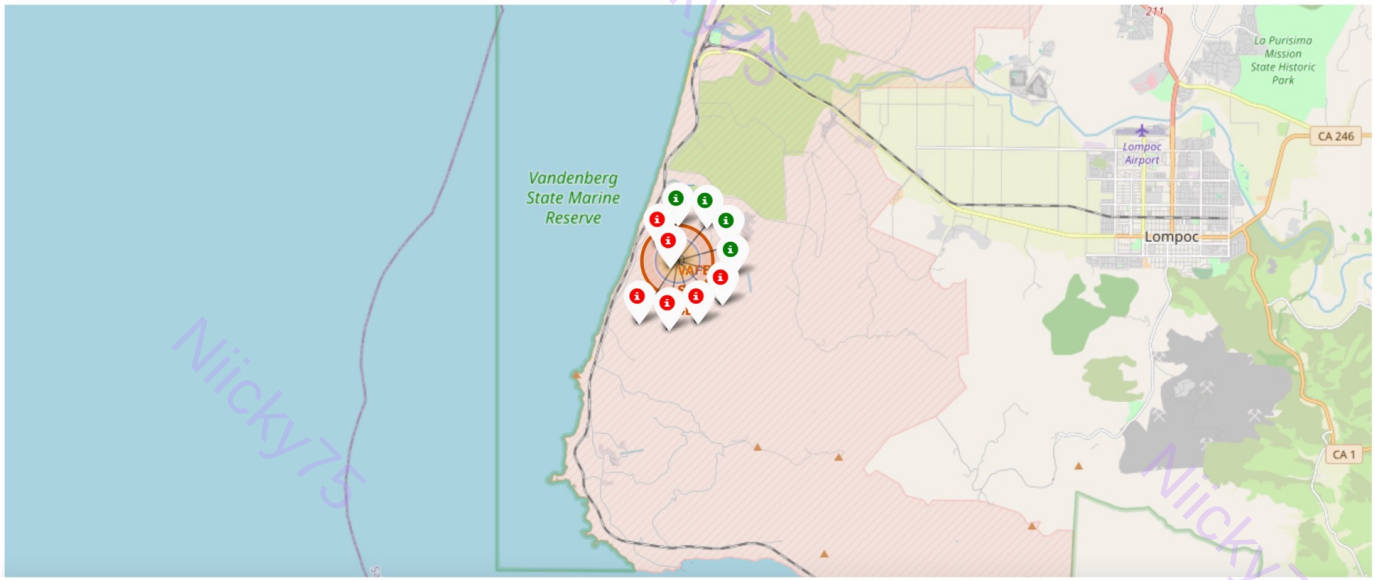
Results: Folium

- Launch sites on the map:



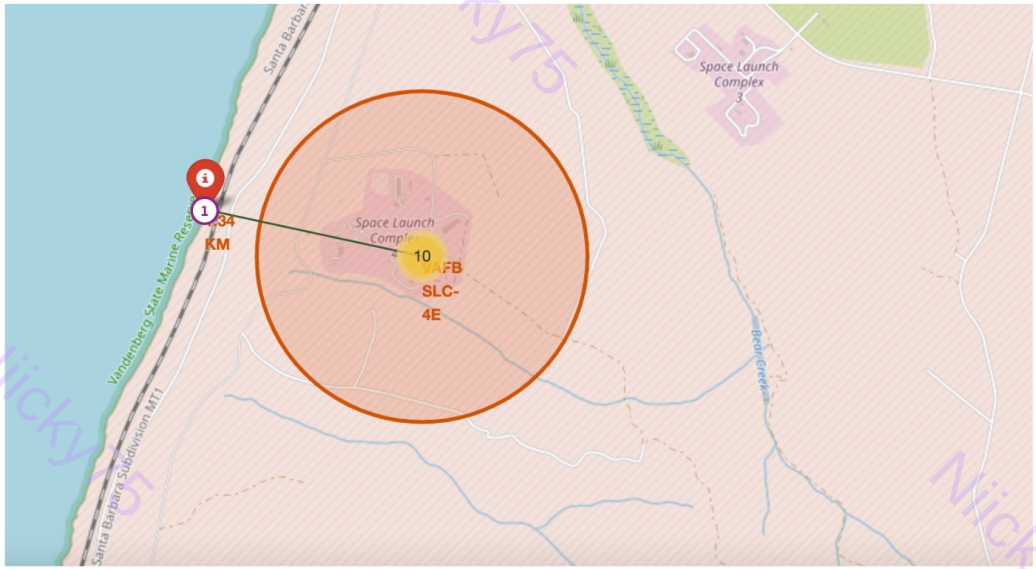
Results: Folium

- Succeeded launches and failed launches for each site on map
  - If we zoom in on one of the launch site, we can see green and red tags. Each green tag represents a successful launch while each red tag represents a failed launch



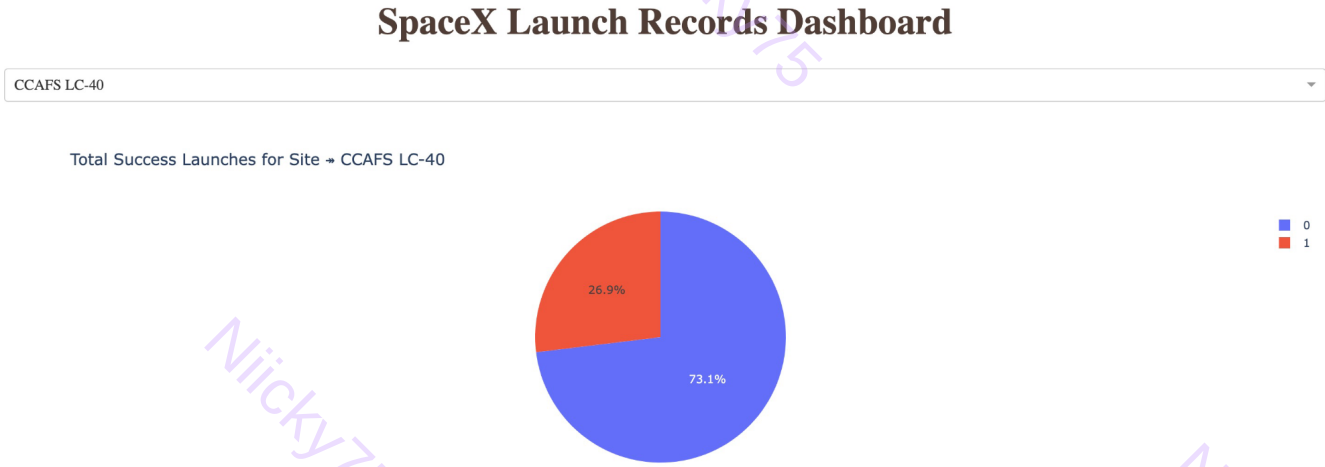
Results: Folium

- Distances between a launch site to its proximities such as the nearest city, railway or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline



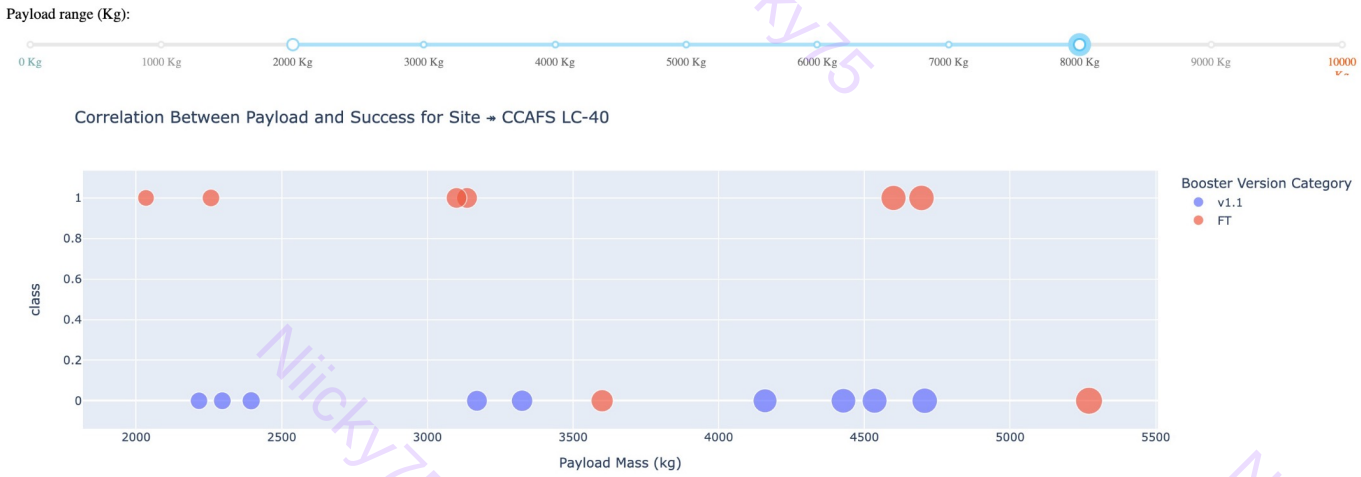
Results : Plotly Dash

- Pie chart when launch site CCAFS LC-40 is chosen.
- 0 represents failed launches while 1 represents successful launches. 73.1% of launches done at CCAFS LC-40 are failed launches.



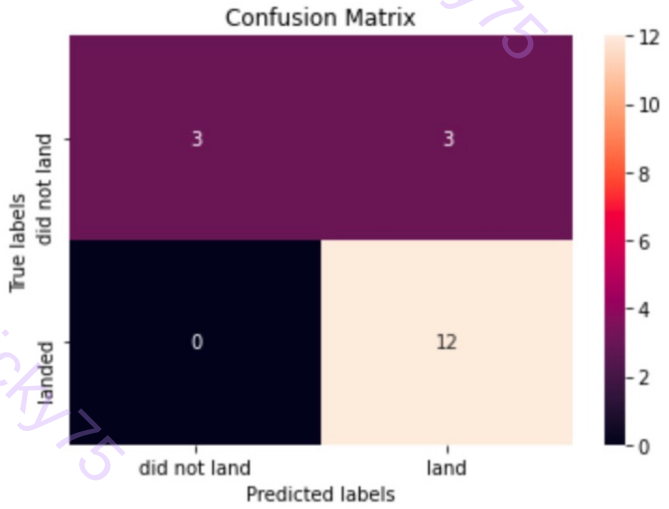
Results : Plotly Dash

- Scatterplot when the payload mass range is set to be
- Class 0 represents failed launches while class 1 represents successful launches.



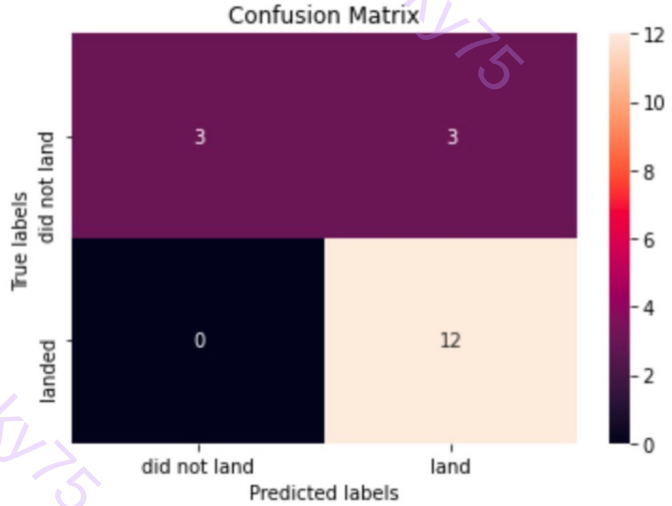
Results: Predictive Analysis

- Logistic regression:
  - GridSearchCV best score: 0.846428571428571 3
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



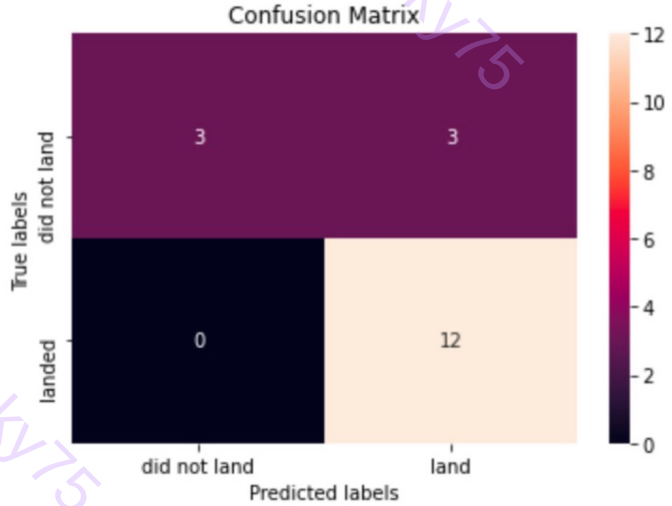
Results: Predictive Analysis

- Support Vector Machine (SVM):
  - GridSearchCV best score: 0.8482142857142856
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



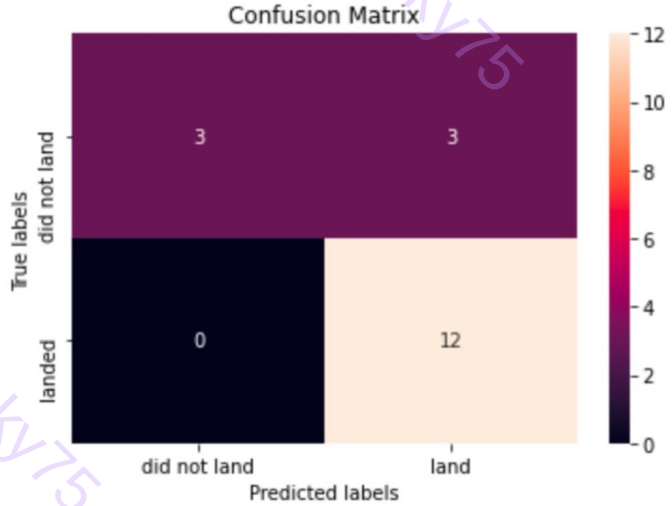
Results: Predictive Analysis

- Decision tree:
  - GridSearchCV best score: 0.8928571428571429
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



Results: Predictive Analysis

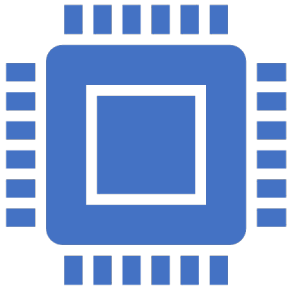
- K-nearest neighbors (KNN):
  - GridSearchCV best score: 0.8482142857142858
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



Results: Predictive Analysis

- Putting the results of all 4 models side by side, we can see that they all share the same accuracy score and confusion matrix when tested on the test set.
- Therefore, their GridSearchCV best scores are used to rank them instead. Based on the GridSearchCV best scores, the models are ranked in the following order with the first being the best and the last one being the worst:
    - Decision tree (GridSearchCV best score: 0.8892857142857142)
    - K nearest neighbors, KNN (GridSearchCV best score: 0.8482142857142858)
    - Support vector machine, SVM (GridSearchCV best score: 0.8482142857142856)
    - Logistic regression (GridSearchCV best score: 0.8464285714285713)

Discussion



From the data visualization, it is evident that certain features might be linked to the mission outcome in various ways. For instance, missions with heavy payloads show a higher success rate for landings in Polar, LEO, and ISS orbits. However, for GTO, the data does not clearly differentiate between successful and unsuccessful landings, as both outcomes are present.

Therefore, each feature might influence the final mission outcome to some extent. Determining the precise impact of each feature is challenging. However, we can utilize machine learning algorithms to analyze patterns in historical data and predict the success of a mission based on these features.

Conclusion

In this project, we aim to predict whether the first stage of a Falcon 9 launch will land successfully, as this is crucial for determining the launch cost.

Each feature of a Falcon 9 launch, such as payload mass or orbit type, may influence the mission outcome.

We employ various machine learning algorithms to analyze patterns in historical Falcon 9 launch data and develop predictive models for mission outcomes.

Among the four machine learning algorithms used, the predictive model generated by the decision tree algorithm showed the best performance.