Winning the Space Race with Data Science

Objectives

- ・The focus of our capstone project is to utilize various machine learning classification algorithms to forecast the successful landing of the SpaceX Falcon 9 first stage.

- ・The primary stages of this endeavor encompass data gathering, cleaning, and structuring, followed by exploratory analysis and interactive visualization of the data.

- ・Through our analysis, we discern correlations between certain aspects of rocket launches and their outcomes, distinguishing between success and failure.

- ・Moreover, our findings suggest that the decision tree algorithm exhibits promise as the optimal method for predicting the successful landing of the Falcon 9 first stage.

Introduction

- ・The primary focus of our capstone project is to develop a predictive model for determining the successful landing of the Falcon 9 first stage. This prediction holds significant implications, particularly in estimating the cost-effectiveness of SpaceX's launches compared to other providers. By scrutinizing various features such as payload mass and launch site, we aim to discern patterns that correlate with successful landings. This analysis will not only provide valuable insights for SpaceX's operations but also serve as useful information for competing companies looking to bid for rocket launch contracts.

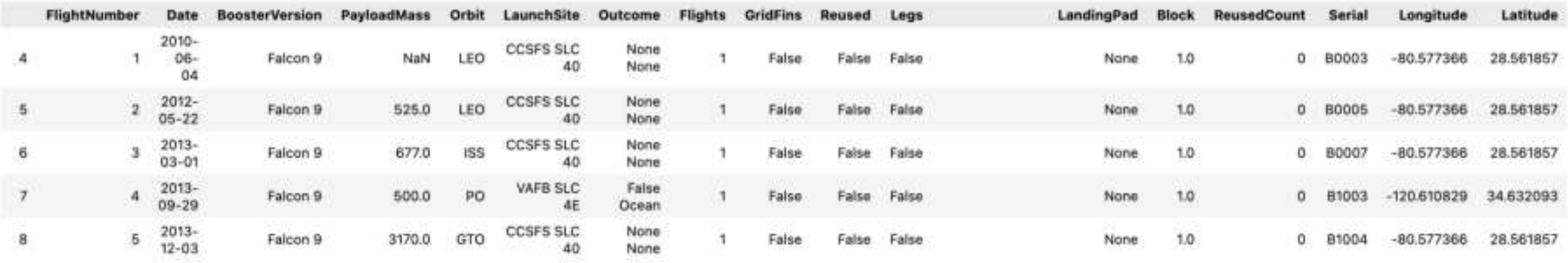
- The overall methodology includes:

1. Data collection, wrangling, and formatting, using: • SpaceX API • Web scraping
2. Exploratory data analysis (EDA), using: • Pandas and NumPy • SQL
3. Data visualization, using: • Matplotlib and Seaborn • Folium • Dash
4. Machine learning prediction, using • Logistic regression • Support vector machine (SVM) • Decision tree • K-nearest neighbors (KNN)

Data Collection

- The SpaceX API utilized for this project can be accessed via https://api.spacexdata.com/v4/rockets/. This API furnishes comprehensive data concerning various rocket launches conducted by SpaceX. To streamline our analysis, we specifically filter the data to encompass only Falcon 9 launches. In the process of preparing the data for analysis, any missing values within the dataset are imputed with the mean of the respective column to ensure completeness and accuracy. After this preprocessing step, our dataset comprises 90 rows or instances, each characterized by 17 columns or features. The initial few rows of the dataset are depicted in the image below for reference.

Data Collection

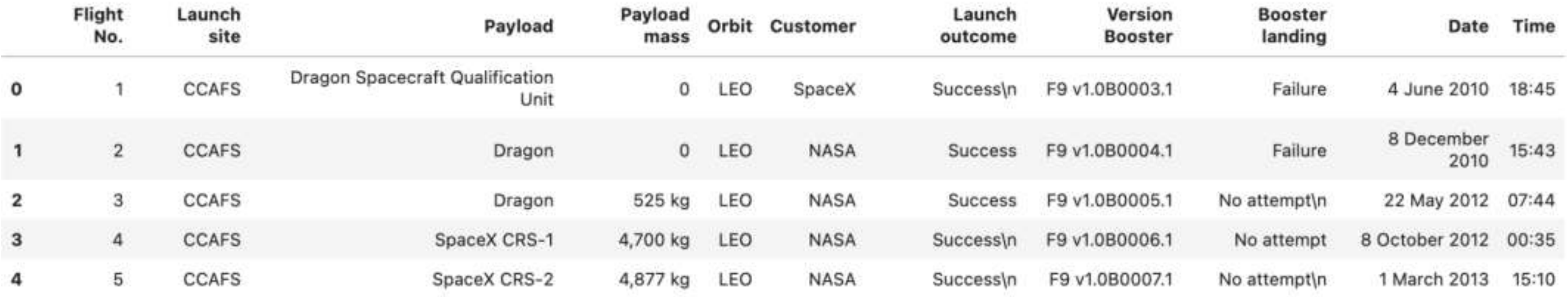


	FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial	Longitude	Latitude
4	1	2010-06-04	Falcon 9	NaN	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003	-80.577366	28.561857
5	2	2012-05-22	Falcon 9	525.0	LEO	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

Data collection

・The data is scraped from https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches& oldid=1027686922

・The website contains only the data about Falcon 9 launches. ・We end up with 121 rows or instances and 11 columns or features. The picture below shows the first few rows of the data:



	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

Data Visualization

- Pandas and NumPy
- Functions from the Pandas and NumPy libraries are used to derive basic information about the data collected, which includes:
- The number of launches on each launch site
- The number of occurrence of each orbit
- The number and occurrence of each mission outcome
- ・SQL・The data is queried using SQL to answer several questions about the data such as:・The names of the unique launch sites in the space mission・The total payload mass carried by boosters launched by NASA (CRS)・The average payload mass carried by booster version F9 v1.1

Data Visualization

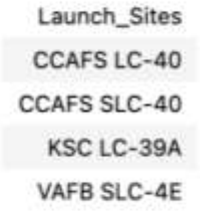
- Matplotlib and Seaborn •
- Functions from the Matplotlib and Seaborn libraries are used to visualize the data through scatterplots, bar charts, and line charts. • The plots and charts are used to understand more about the relationships between several features, such as:
- • The relationship between flight number and launch site • The relationship between payload mass and launch site • The relationship between success rate and orbit type
- Folium
- • Functions from the Folium libraries are used to visualize the data through interactive maps.
- The Folium library is used to: • Mark all launch sites on a map • Mark the succeeded launches and failed launches for each site on the map • Mark the distances between a launch site to its proximities such as the nearest city, railway, or highway

Machine Learning

- For our machine learning modeling, we leverage functions from the Scikit-learn library. The prediction phase comprises several key steps:
- Standardizing the data
- Splitting the data into training and test sets
- Creating machine learning models, including:
  - Logistic regression
  - Support vector machine (SVM)
  - Decision tree
  - K nearest neighbors (KNN)
- Fitting the models on the training set
- Identifying the optimal combination of hyperparameters for each model
- Evaluating the models based on their accuracy scores and confusion matrices.

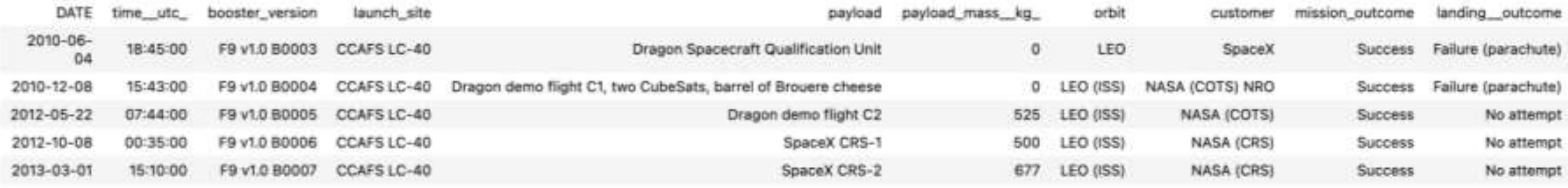
Results

- The names of the unique launch sites in the space mission



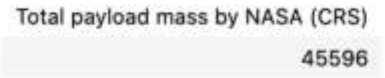
Launch_Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

- 5 records where launch sites begin with 'CCA'



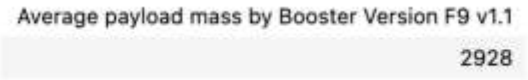
DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight 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

- The total payload mass carried by boosters launched by NASA (CRS)



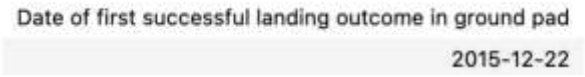
Total payload mass by NASA (CRS)
45596

- The average payload mass carried by booster version F9 v1.1



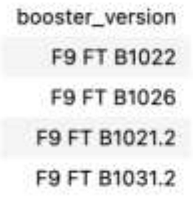
Average payload mass by Booster Version F9 v1.1
2928

- The date when the first successful landing outcome in ground pad was achieved



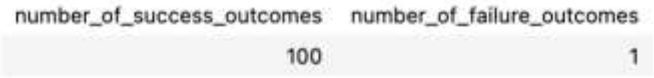
Date of first successful landing outcome in ground pad
2015-12-22

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



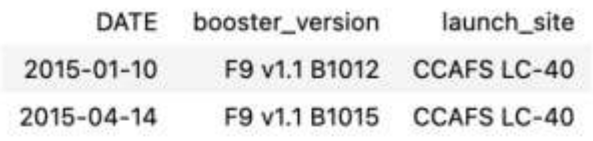
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The total number of successful and failure mission outcomes



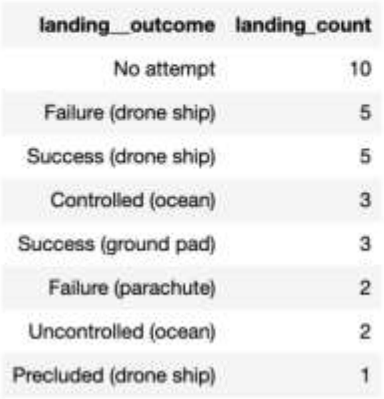
number_of_success_outcomes	number_of_failure_outcomes
100	1

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015



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

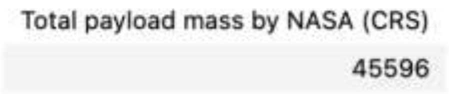
- The count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order



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

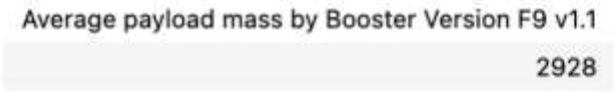
Results (SQL)

- The total payload mass carried by boosters launched by NASA (CRS)



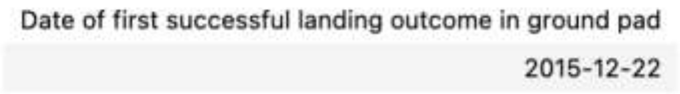
Total payload mass by NASA (CRS)
45596

- The average payload mass carried by booster version F9 v1.1



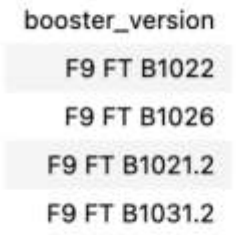
Average payload mass by Booster Version F9 v1.1
2928

- The date when the first successful landing outcome in ground pad was achieved



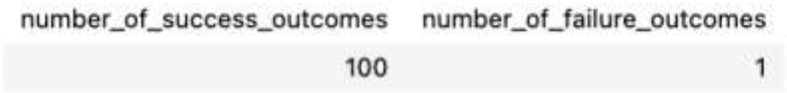
Date of first successful landing outcome in ground pad
2015-12-22

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



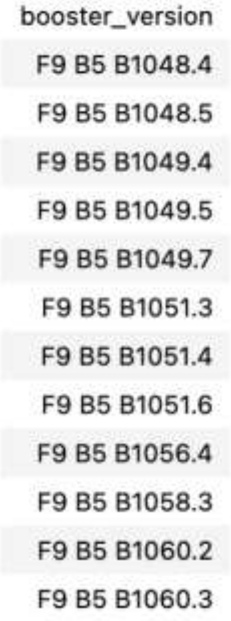
booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

- The total number of successful and failure mission outcomes



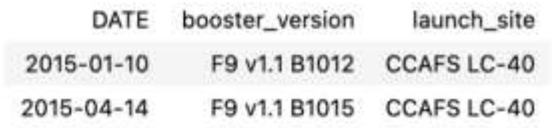
number_of_success_outcomes	number_of_failure_outcomes
100	1

- The names of the booster versions which have carried the maximum payload mass



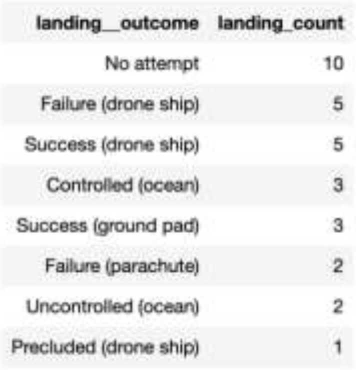
booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

- The failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015



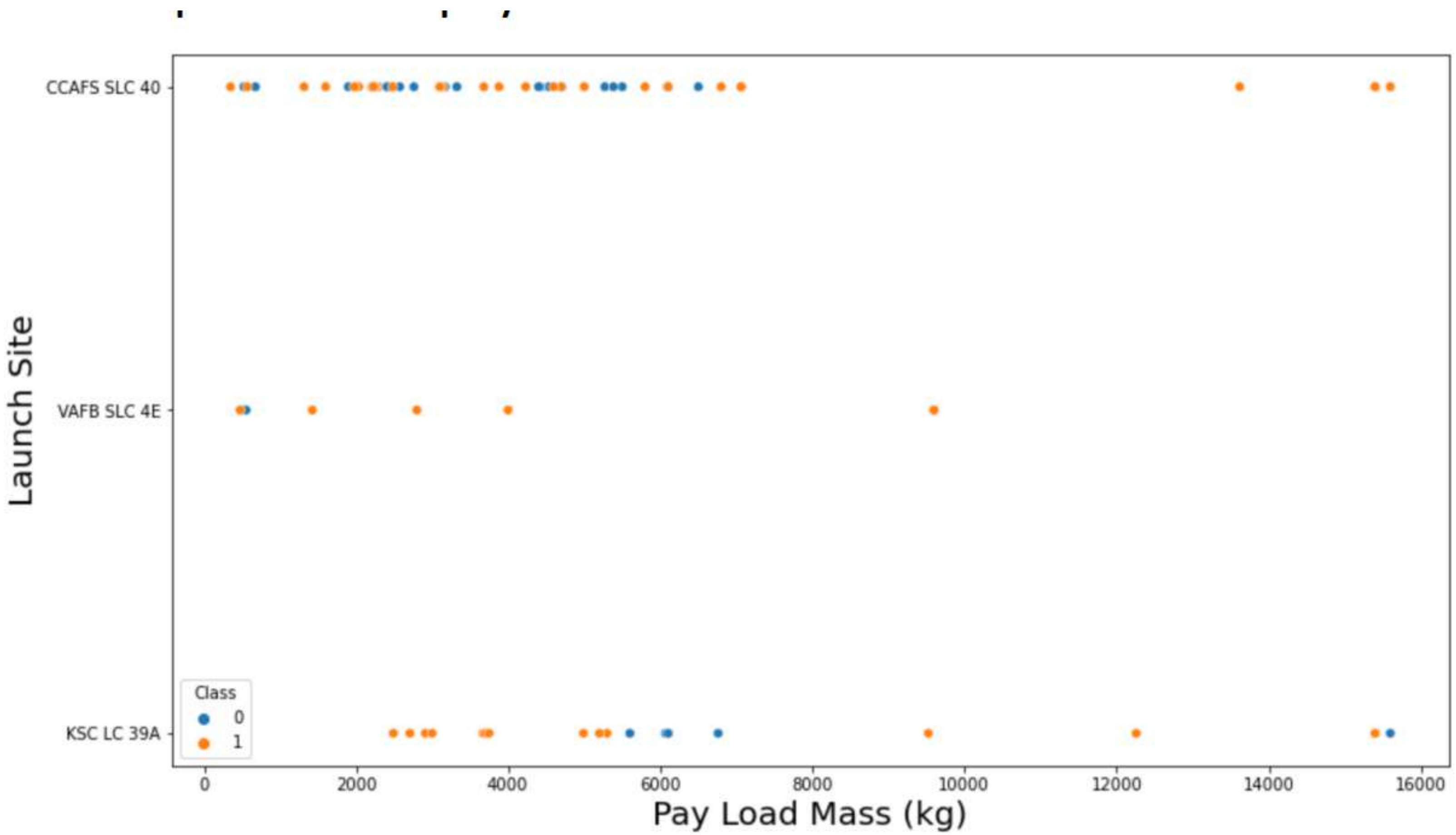
DATE	booster_version	launch_site
2015-01-10	F9 v1.1 B1012	CCAFS LC-40
2015-04-14	F9 v1.1 B1015	CCAFS LC-40

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

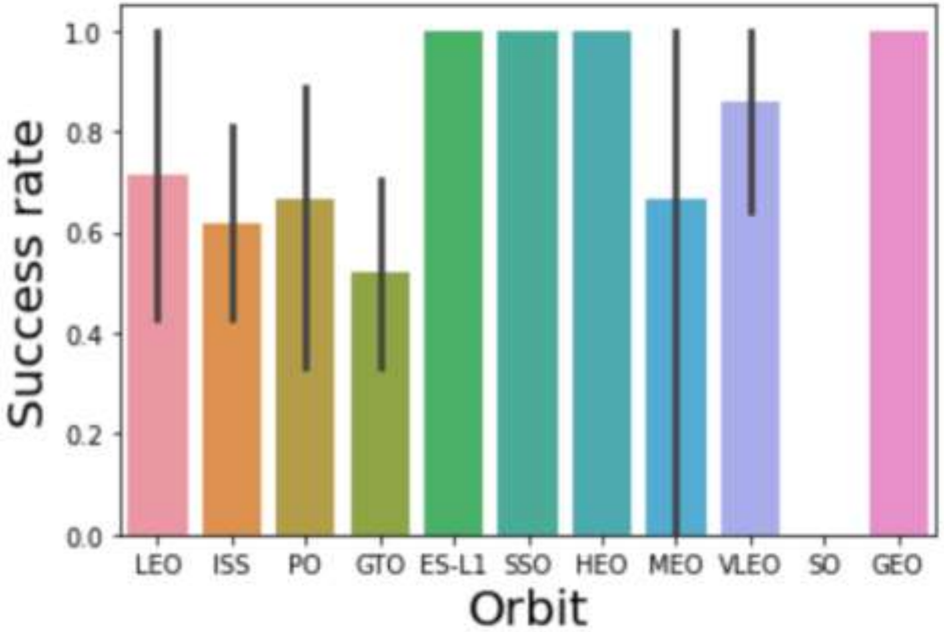


landing__outcome	landing_count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

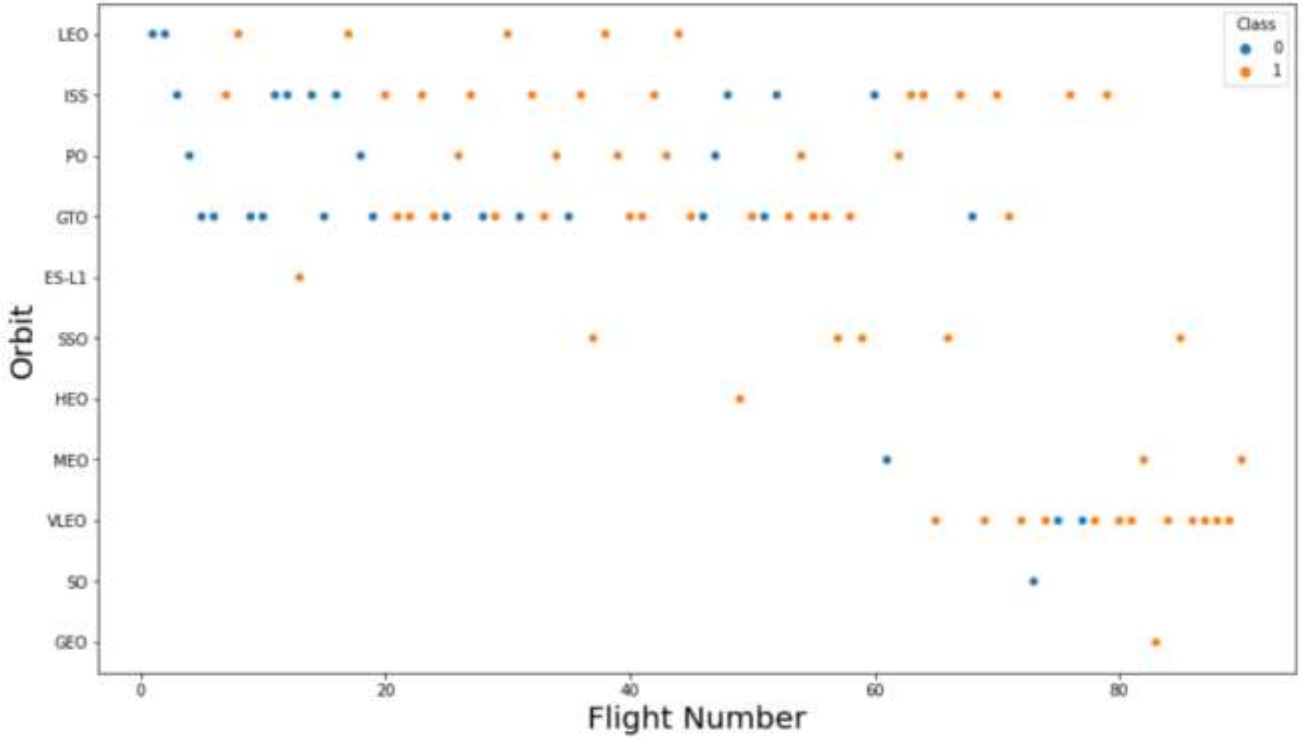
Results (Plots)



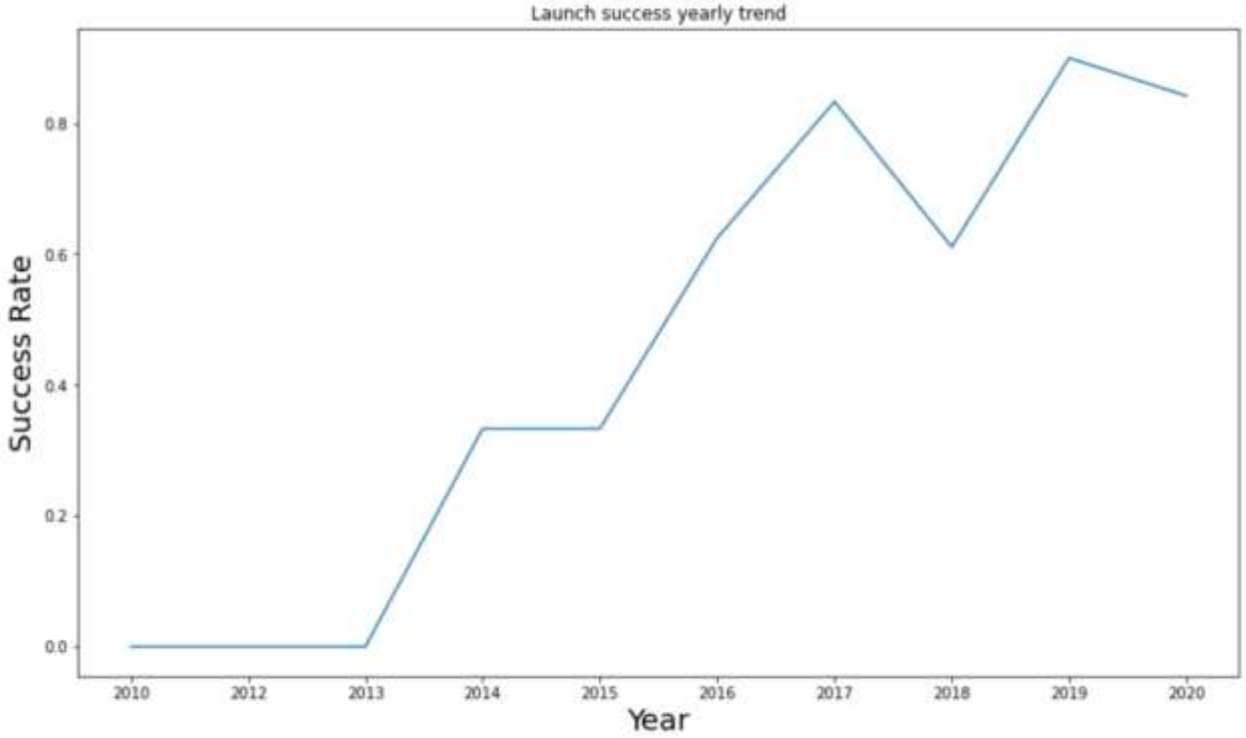
- The relationship between success rate and orbit type



- The relationship between flight number and orbit type

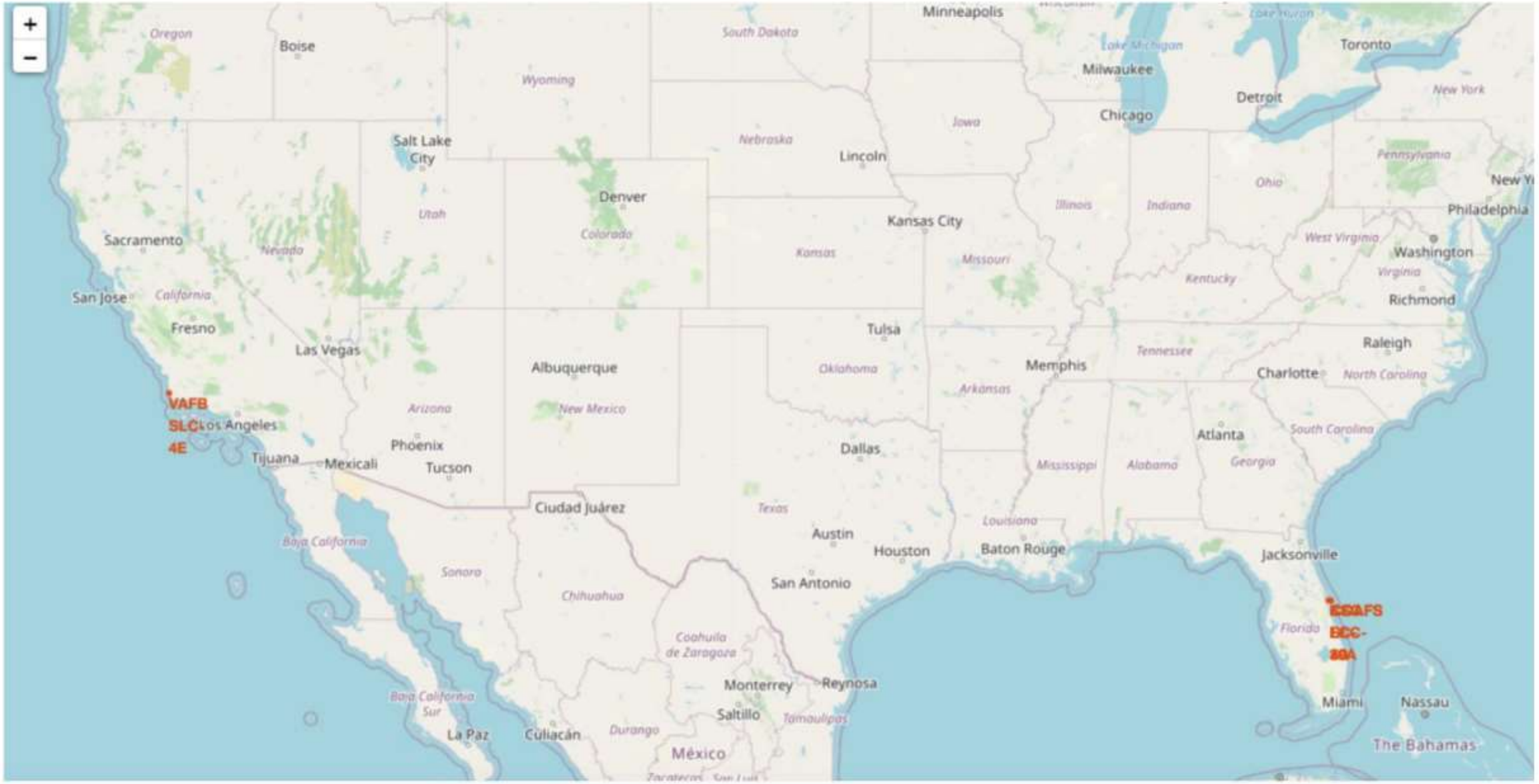


- The launch success yearly trend

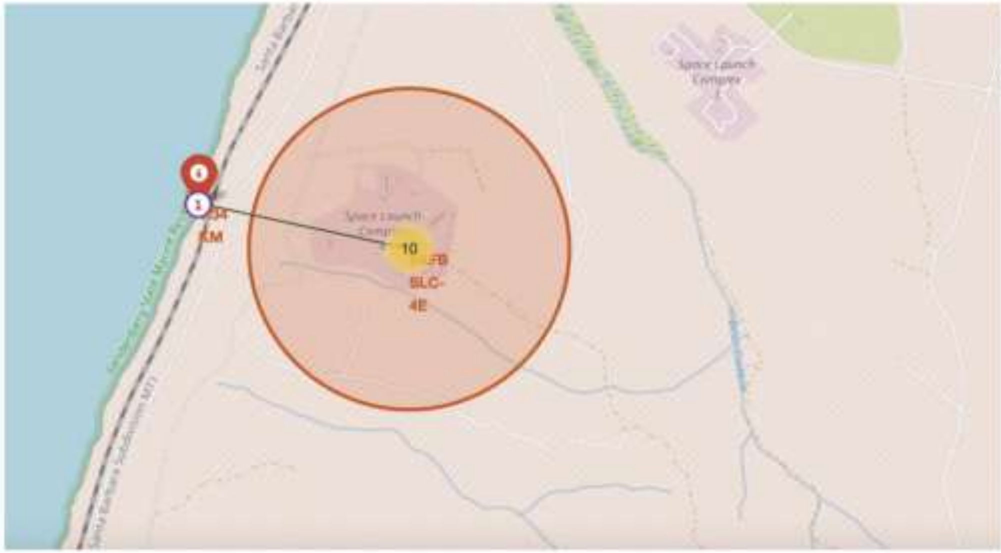


Results (folium)

- Launch sites:

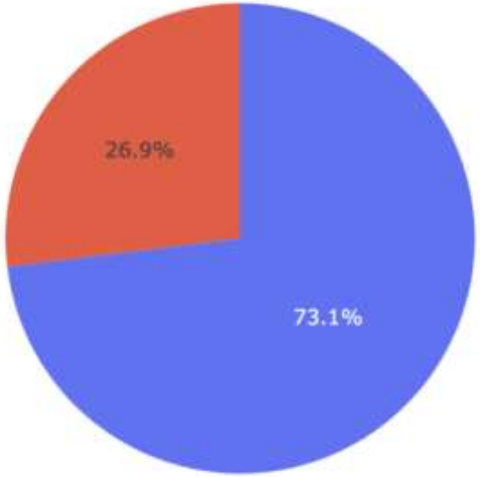


- The distances between a launch site to its proximities such as the nearest city, railway, or highway
  - The picture below shows the distance between the VAFB SLC-4E launch site and the nearest coastline

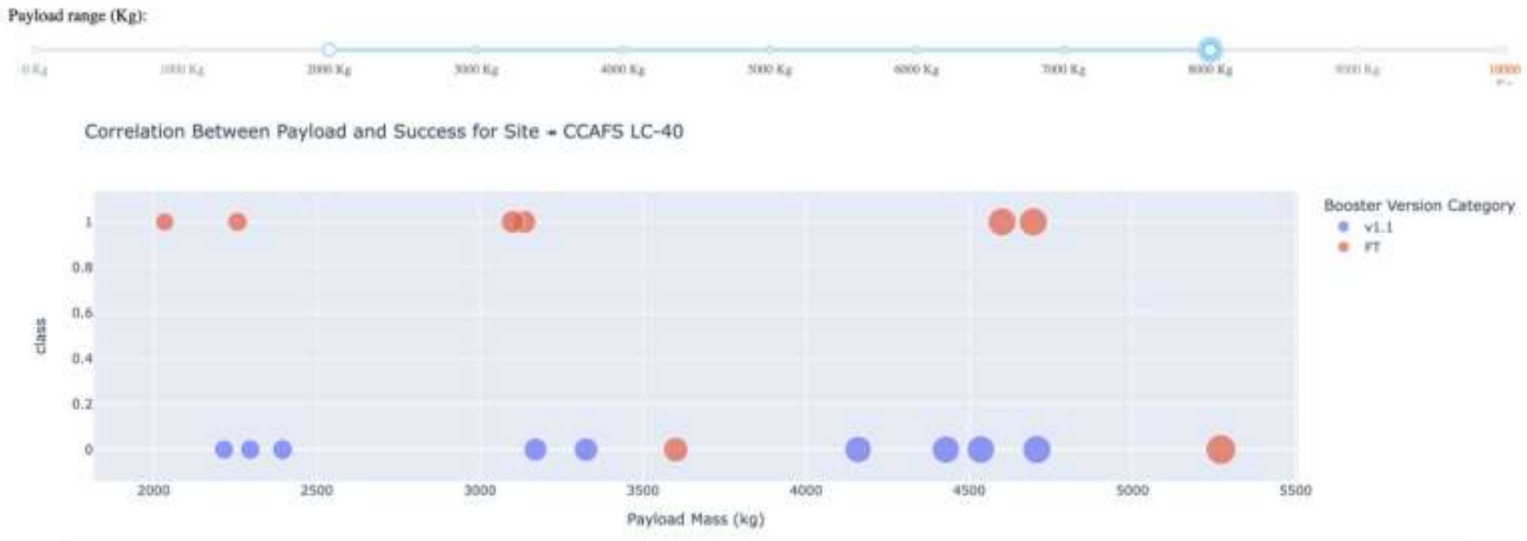


Results (dash)

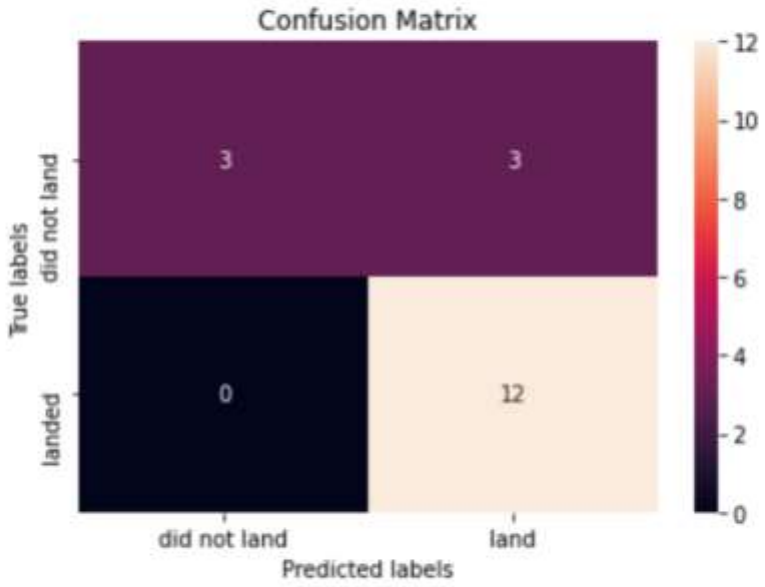
- The pie chart displayed below represents the outcome of launches specifically from launch site CCAFS LC-40. In this chart, a value of 0 indicates failed launches, while 1 represents successful launches. Notably, the data illustrates that approximately 73.1% of launches conducted at CCAFS LC-40 have resulted in failures.



- The picture below shows a scatterplot when the payload mass range is set to be from 2000kg to 8000kg.
- Class 0 represents failed launches while class 1 represents successful launches.



- Logistic regression
  - GridSearchCV best score: 0.8464285714285713
  - Accuracy score on test set: 0.8333333333333334
  - Confusion matrix:



Results (predctive analysis)

- Upon comparing the results of all four models, it's evident that they share identical accuracy scores and confusion matrices when evaluated on the test set. Consequently, we resort to their GridSearchCV best scores to rank them. Based on these scores, the models are ranked as follows, with the first being the most favorable and the last being the least favorable:
- 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

-

The data visualization section reveals potential correlations between certain features and mission outcomes. For instance, heavy payloads exhibit higher rates of successful landings or positive outcomes, particularly for orbit types Polar, LEO, and ISS. However, distinguishing such trends becomes more challenging for orbit type GTO, where both positive and negative outcomes occur. Hence, each feature likely influences the final mission outcome to some extent. While the precise impact of these features remains elusive, employing machine learning algorithms enables us to discern patterns from historical data and predict the likelihood of a mission's success based on the provided features.

Conclusion

- In this project, our aim is to predict whether the first stage of a given Falcon 9 launch will successfully land, with the ultimate goal of estimating the launch cost. We recognize that various features associated with a Falcon 9 launch, such as payload mass or orbit type, may influence the mission outcome in distinct ways. To achieve this prediction, we employ several machine learning algorithms to discern patterns within past Falcon 9 launch data, thus generating predictive models capable of forecasting the outcome of future launches. Among the four machine learning algorithms utilized, the decision tree algorithm emerges as the top performer, exhibiting superior predictive capabilities compared to the others.