IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rafael Basso

22/07/2022

Outline

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

Executive Summary

- Methodologies Summary
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Visualization
  - Exploratory Data Analysis with SQL
  - Interactive Map Building
  - Dashboard Creation
  - Predictive Analysis
- Results Summary
  - Exploratory Data Analysis Results
  - Classification Model Building

Introduction

- Project background and context
  - SpaceX is an American spacecraft manufacturer, space launch provider, and a satellite communications corporation founded in 2002 by Elon Musk, with the goal of reducing space transportation costs to enable the colonization of Mars. SpaceX manufactures the Falcon 9 and Falcon Heavy launch vehicles, several rocket engines, Cargo Dragon, crew spacecraft, and Starlink communications satellites.
  - SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is due to the fact that SpaceX can reuse the first stage. So, the best way to reduce the cost is to determine whether the first stage will land or not.
- Problems you want to find answers
  - Is there a way to use machine learning and predict with good accuracy if the first stage will land safely?
  - What features have more impact on a successful landing?
  - Which features are determinant to cost reduction of a launch?

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology
  - Usage of two public data sources:
    - Space X REST API
    - WebScraping from Wikipedia
- Perform data wrangling
  - Filter data
  - Missing data dealing
  - Binary Classification creation, representing success or failure of landing

Methodology

Executive Summary

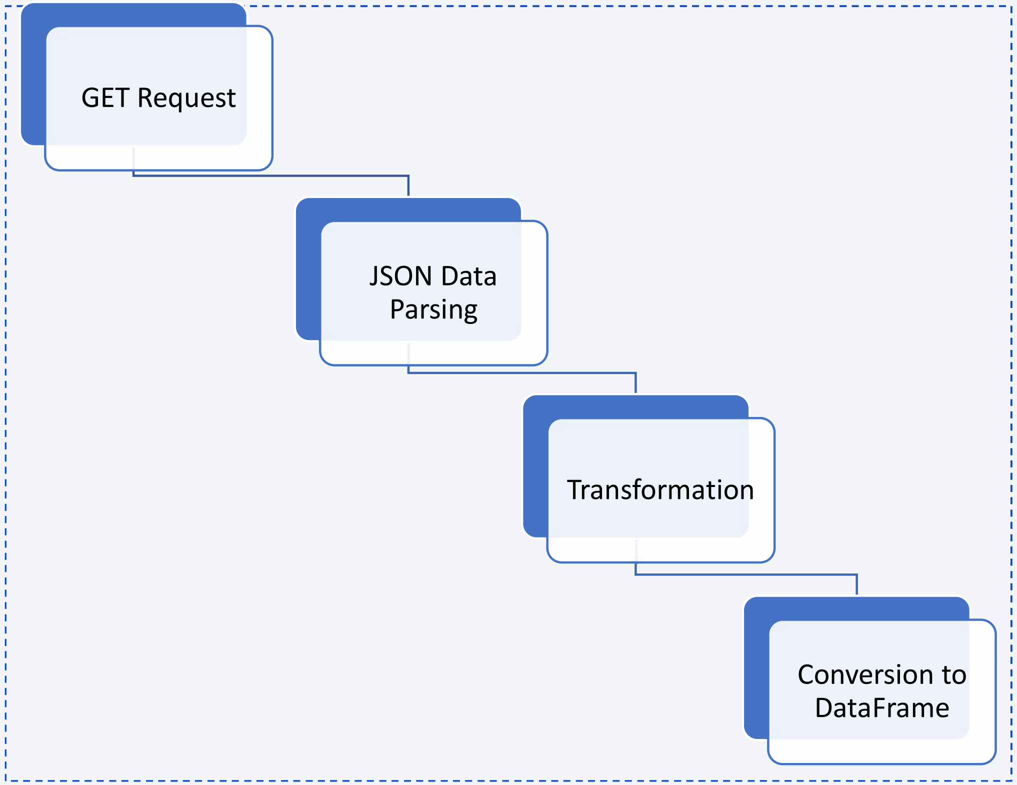
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Creation, tunning and evaluation of different classification models to predict the result of the landing
  - Evaluation of different accuracy methods

Data Collection

- Data Collection process involved the usage of two different techniques:
  - Requests from SpaceX REST API
  - Webscraping the SpaceX Wikipedia website

Data Collection – SpaceX API

- GET Request to SpaceX API
- Data converted from Json
- Column filters and feature extraction
- Data converted to DataFrame

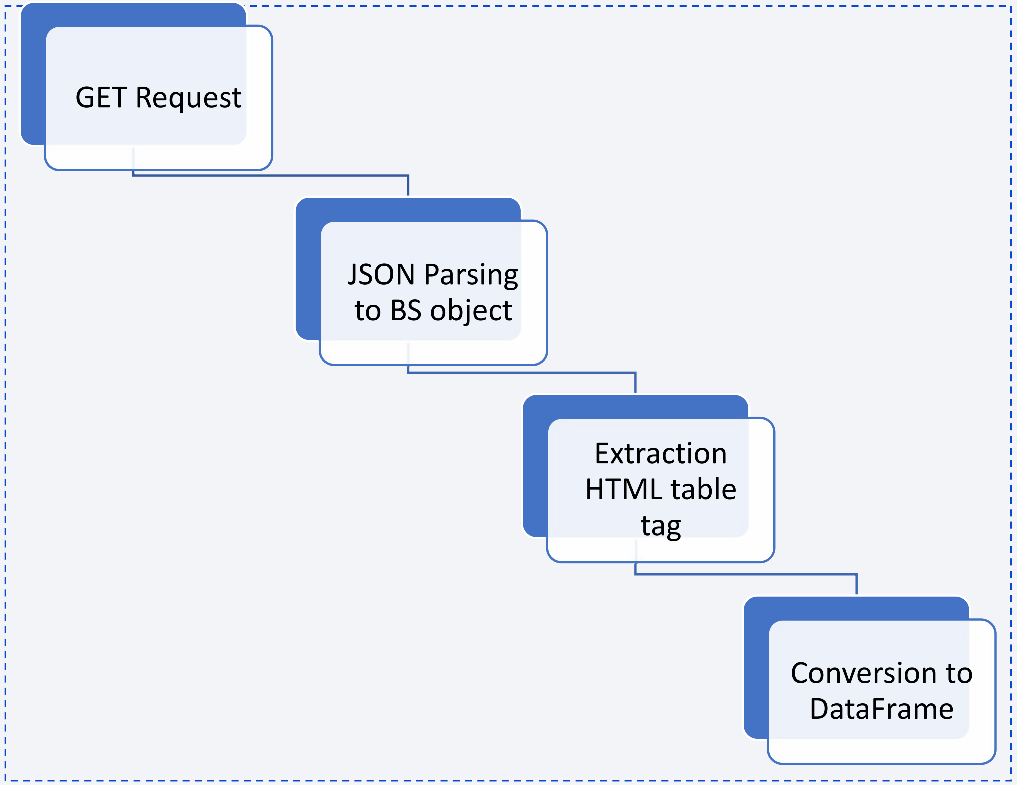


[Rafael Basso Github Shared Notebook]

Data Collection – Scraping

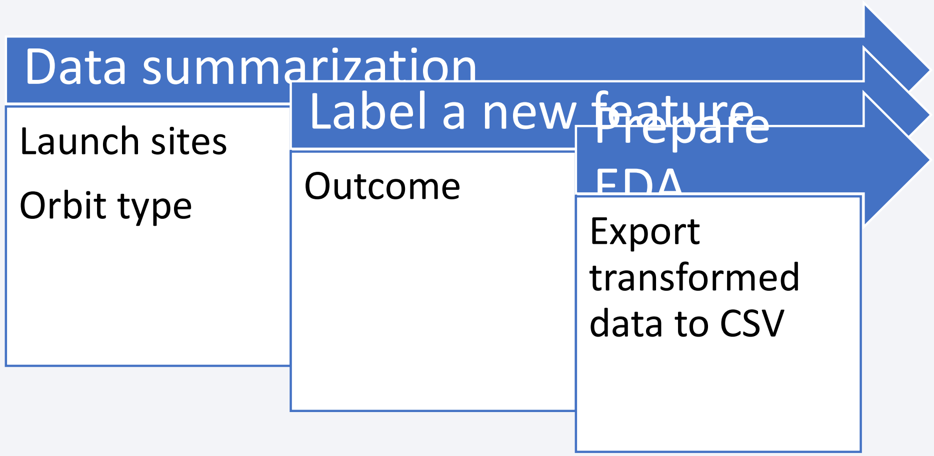
- GET Request to SpaceX Wikipedia
- Data converted from Json to BeautifulSoup object
- Extract HTML table tags
- Data converted to DataFrame

[Rafael Basso Github Shared Notebook]



Data Wrangling

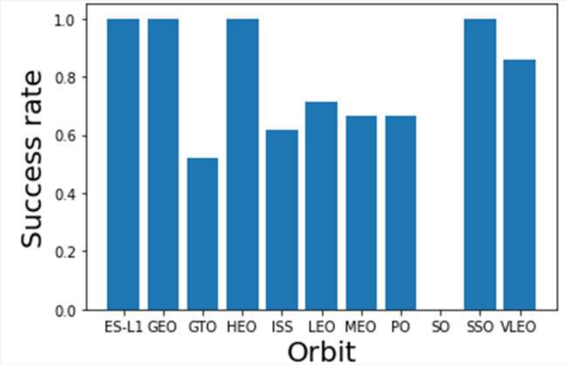
- CSV Reading
- Missing Data Dealing
- Exploratory Data Analysis
  - Summarization
  - Number of launches per site and orbit type
- Creation of Label Feature Outcome
- Data Export to CSV



[Rafael Basso Github Shared Notebook]

EDA with Data Visualization

- Usage of Scatterplots to visualize the relationship between the features
  - Payload Mass X Flight Number, Launch Site X Flight Number, Launch Site X Payload Mass, Orbit and Flight Number, Payload and Orbit
- Barplots were used to evaluate the influence of Orbit feature in the success rate



Rafael Basso Github Shared Notebook

EDA with SQL

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was acheived.
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
- List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
- 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

[Rafael Basso Github Shared Notebook]()

Build an Interactive Map with Folium

- Markers, circles, lines and marker clusters were used with Folium Maps
  - Markers show launch sites;
  - Circles show highlighted areas around specific coordinates (e.g. NASA Johnson Space Center)
  - Marker clusters show different launch positions in a same launch site
  - Lines show distances between two coordinates.

Rafael Basso Github Shared Notebook

Build a Dashboard with Plotly Dash

- A dashboard was developed to visualize:
  - Percentage of launches by site
  - Payload range

- Both graphs are suitable to show the relation between payloads and launch sites
  - It is possible to detect what are the best places to launch according to payload range

[Rafael Basso Github Shared Notebook]

Predictive Analysis (Classification)

- Split of Dataset in 2: training and test sets
- 4 different classifiers were tested:

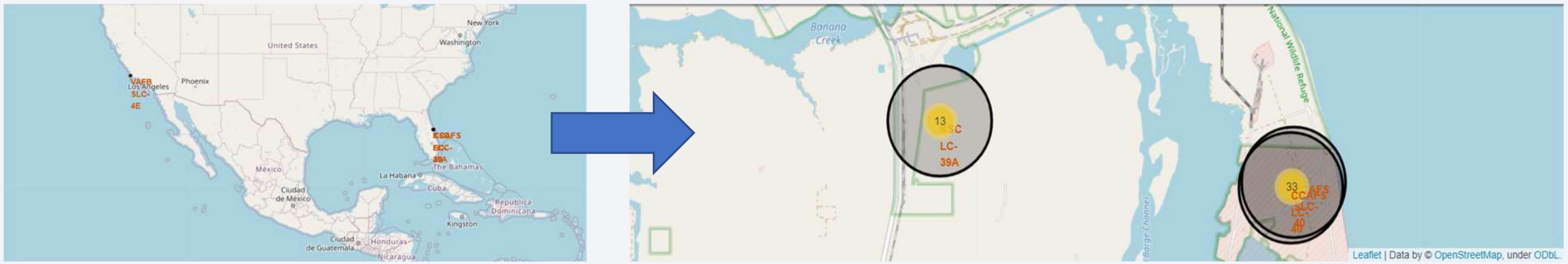
  - Logistic Regression
  - SVM
  - Decision Tree
  - KNN

- Helpful libraries such as sklearn.model_selection and GridSearchCV were used to find out the best hyperparameters for each classifier in the training dataset
- Performance results for each classifier were compared in the test dataset

Rafael Basso Github Shared Notebook

Results

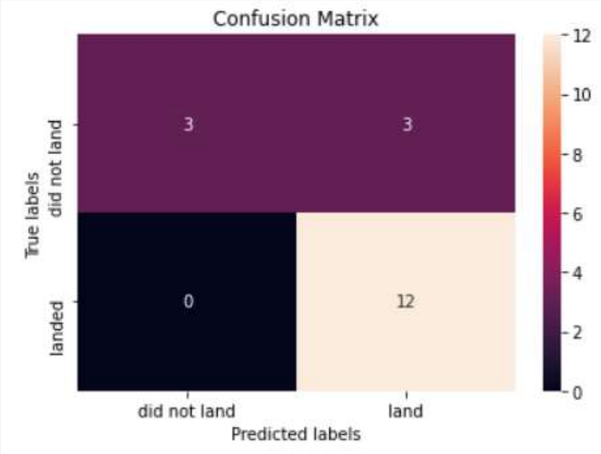
- Exploratory data analysis results
  - Space X uses 4 different launch sites;
  - The average payload of F9 v1.1 booster is 2,928 kg;
  - Falcon 9 booster versions in drone ships tend to have success at landing
  - High successful rates of mission outcomes (almost 100%)
  - Increasing rates of landing outcomes over time
- Exploratory data analysis using geographic data
  - Launch sites are kept in safe places and near coast
  - Most launches happens at east sites (46 launches)



Results

- Predictive Analysis
  - All 4 models had the same accuracy on the test set

	LogisticRegression	SVM	Decision Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333



- Possible Explanation
  - The result can possibly be explained because the test set is too small, with only 18 rows
  - This conclusion is reinforced when we used the models to measure the accuracies with the full dataset and obtained the best accuracy with the SVM model (~87%)

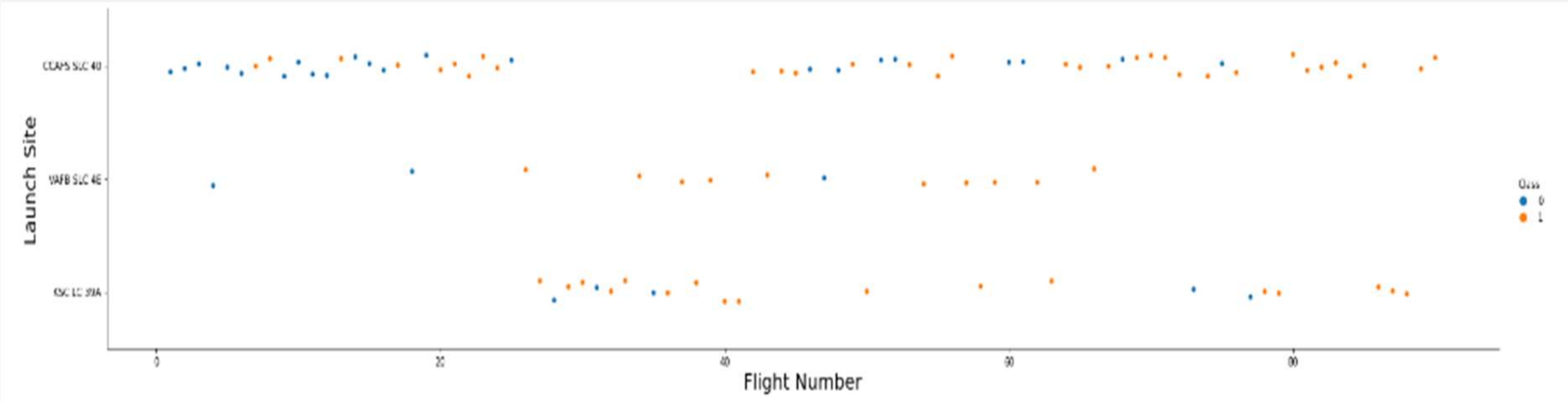
Section 2

Insights drawn from EDA

Flight Number vs. Launch Site

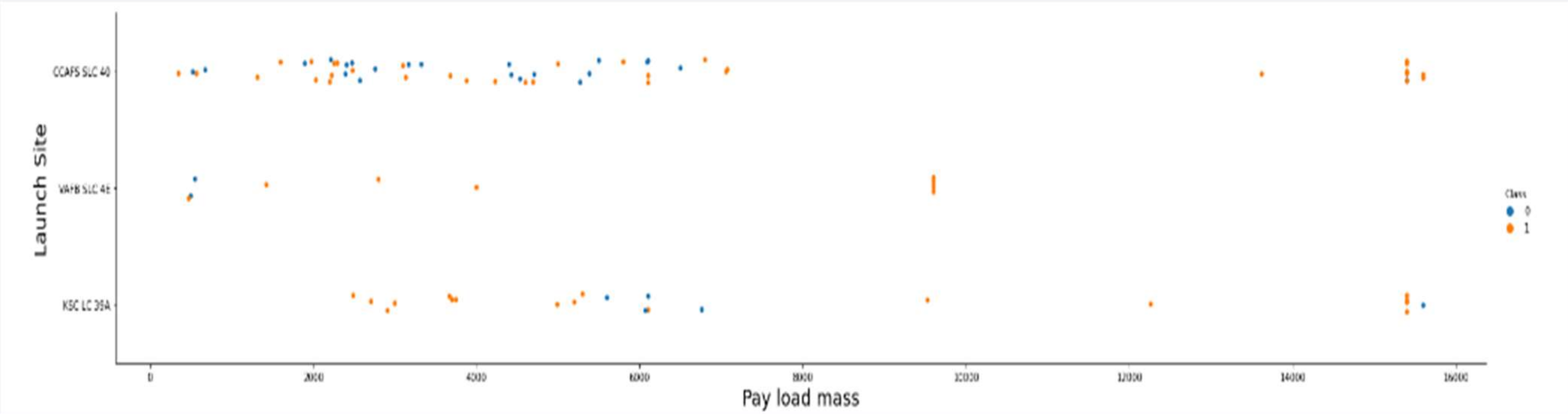
- The launch site with most occurrences is CCAF5 SLC 40, where most of recent launches were successful;

- Success rate improves over time for all launch sites



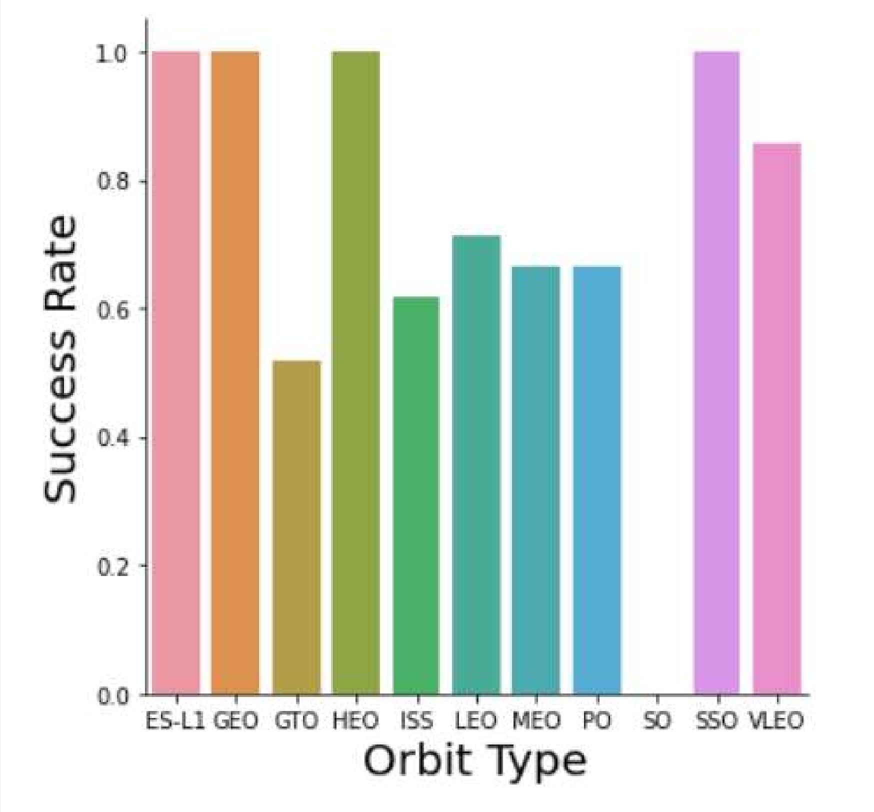
Payload vs. Launch Site

- Payload have a big influence on the success rate
- Most of payloads over 10,000kg were successful
- There is no attempt on VAFB SLC 4E launch site for payload over 10,000kg



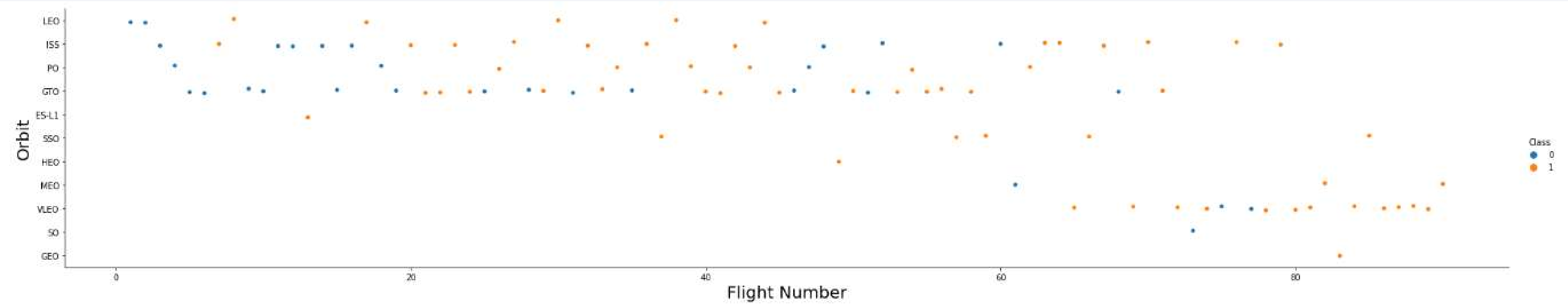
Success Rate vs. Orbit Type

- A few orbits have 100% of success rate, like GEO and HEO
- GTO have the worst success rate



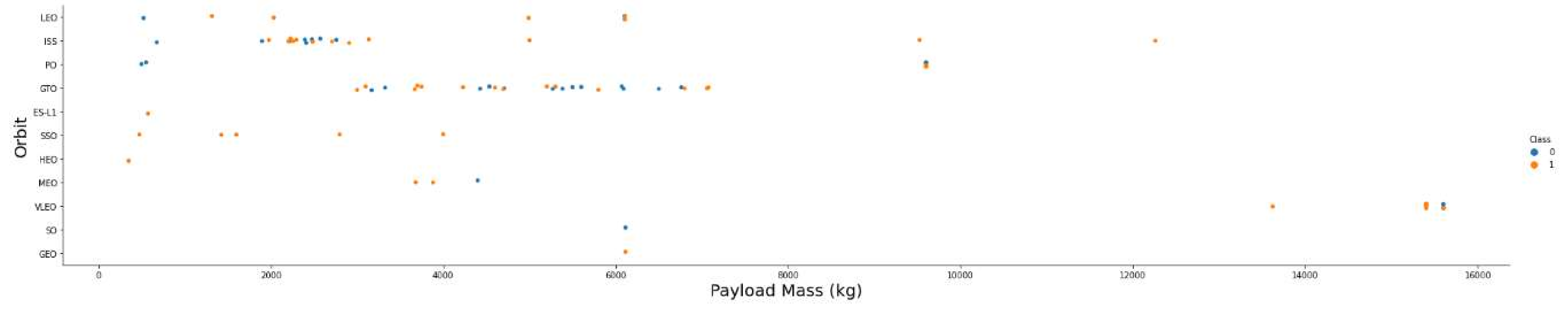
Flight Number vs. Orbit Type

- As the time goes by, the success rate is being improved for all orbits
  - The most recent flights have a better successful rate
- VLEO orbit has been used more recently.



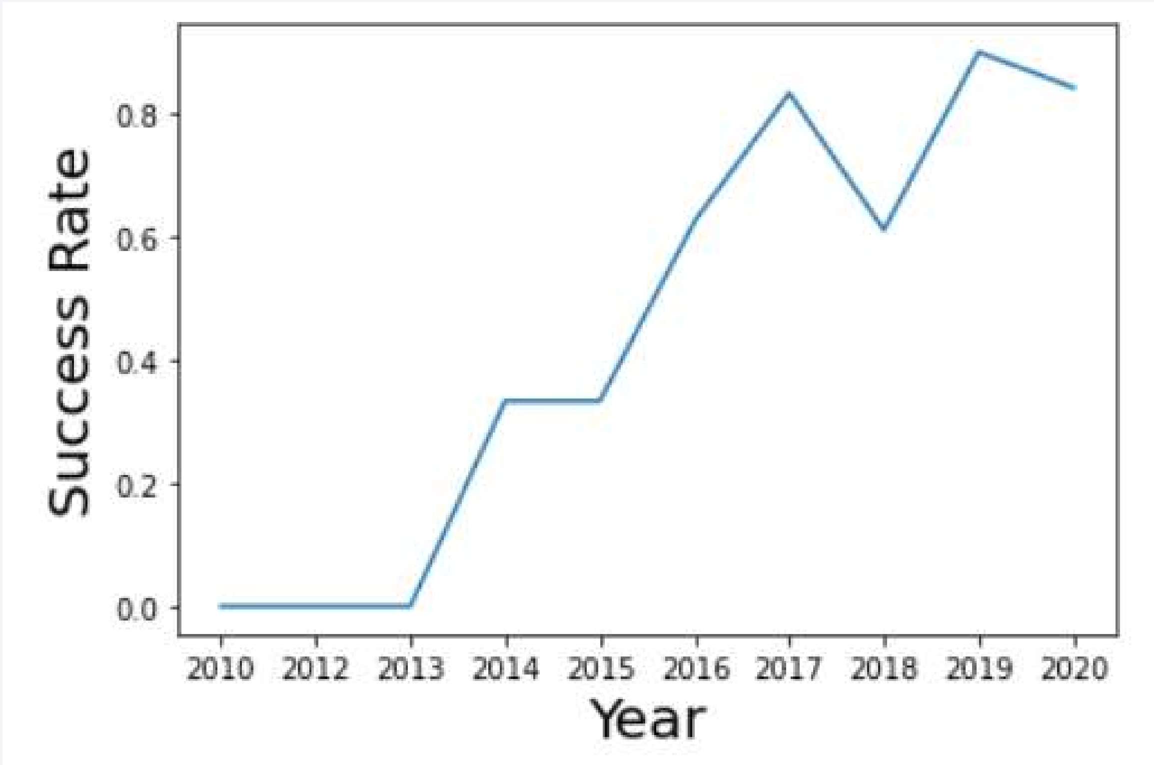
Payload vs. Orbit Type

- Payload and Orbit Type does not seem to have a good correlation
- ISS orbit has the widest range of payload and a good rate of success
- There are just a few attempts to the orbits SO and GEO



Launch Success Yearly Trend

- 2013 is the turning point on the success rate, and apparently the period before was used to develop and improve the technology, because there is no success.
- Since 2013 the success rate kept increasing
- Since 2016 the success rate have always been above 60%



All Launch Site Names

- There are four distinct launch sites, that are result of selecting distinct records from the table SPACEXTBL that was created

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

Launch Site Names Begin with 'CCA'

- We selected all from the SPACEXTBL where launch sites begin with `CCA` and limited the result to 5

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

Total Payload Mass

- A new column named as “total_payload_mass” was created, with the sum of all values in “payload_mass__kg_” column, where the customer equals “NASA (CRS)”

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- The “avg” aggregation function was used to calculate the payload average whose booster version is F9 v1.1

average_payload_mass
2534

First Successful Ground Landing Date

- The “min” aggregation function was used to calculate the first successful landing outcome on ground pad whose outcome value was “Success (ground pad)”

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- To get the needed result, it was used two filters:
  - “Success (drone ship)” on the column “landing__outcome”
  - “payload_mass__kg_” between 4000 and 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

- We grouped all data by "mission_outcomes" and counted all occurrences for each group

mission_outcome	total_number
Failure (in flight)	1
Success	99
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- A subquery was used to get the maximum payload mass. After that, we used the result of this subquery to filter all data which matches with this value

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

- Two filters were used to get the result:
  - “Failure (drone ship)” on the column “landing__outcome”
  - 2015 on the column “date” => This was made using the function “year” on the column, and it was possible because the type of it was DATE

MONTH	DATE	booster_version	launch_site	landing__outcome
January	2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
April	2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

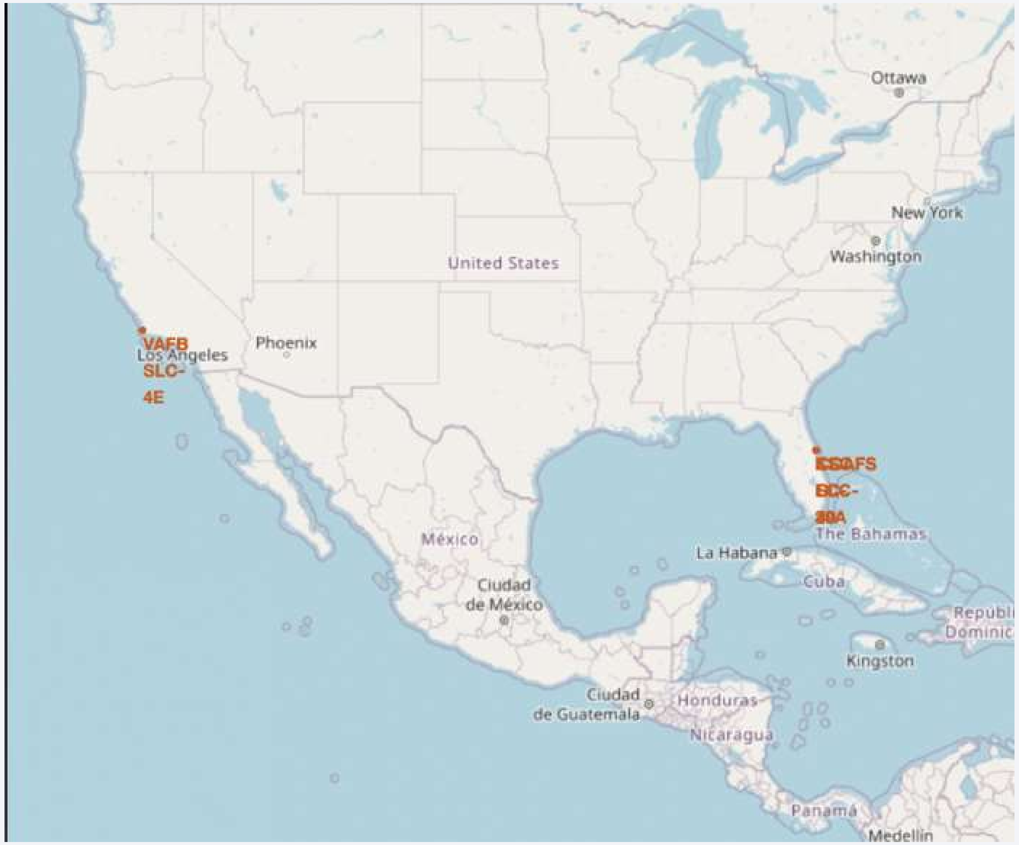
- We used date functions to specify date range, grouped the results by landing outcomes, counted all occurrences for each group values, and showed the results in a descending order

landing__outcome	count_outcomes
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


Section 3
Launch Sites Proximities Analysis

All the launch sites locations

For safety, all the sites are in very close proximity to the coast, while launching rockets towards the ocean it minimizes the risk of having any debris dropping or exploding near people.



Lauch Outcomes by Site

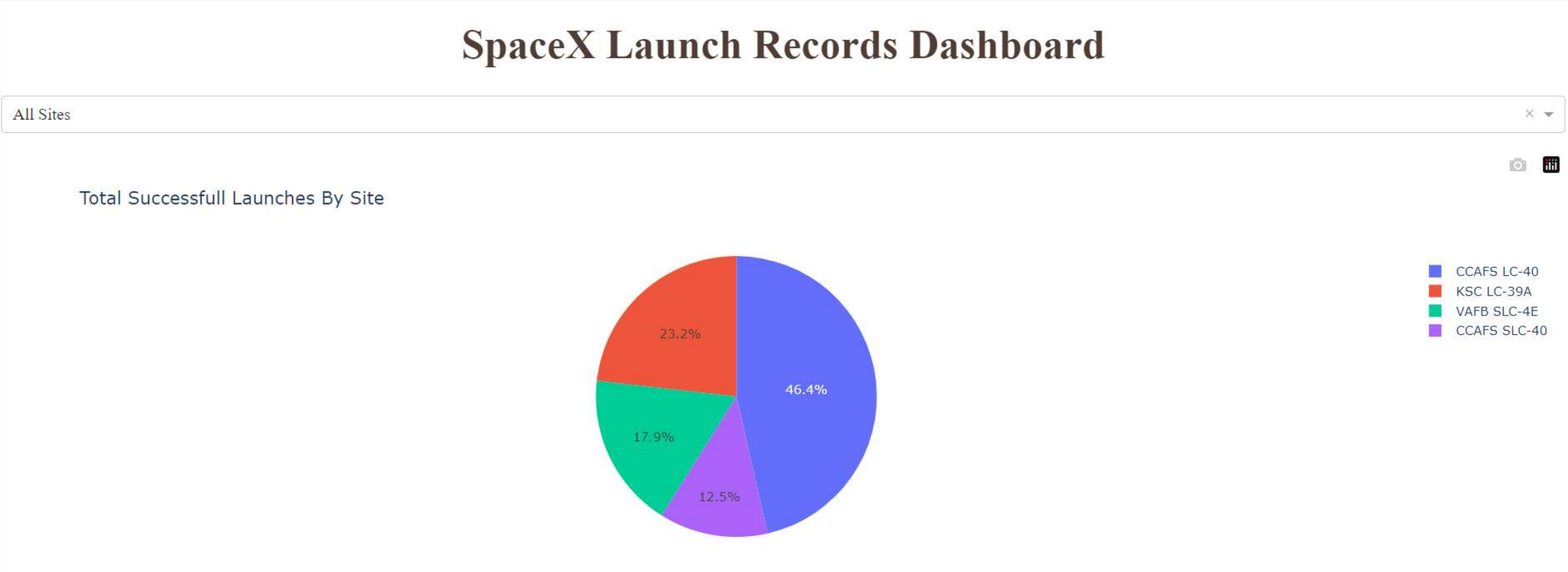
Green marks:
successful launching
Red marks:
failures

Suitable place: near roads and no habitants

Section 4

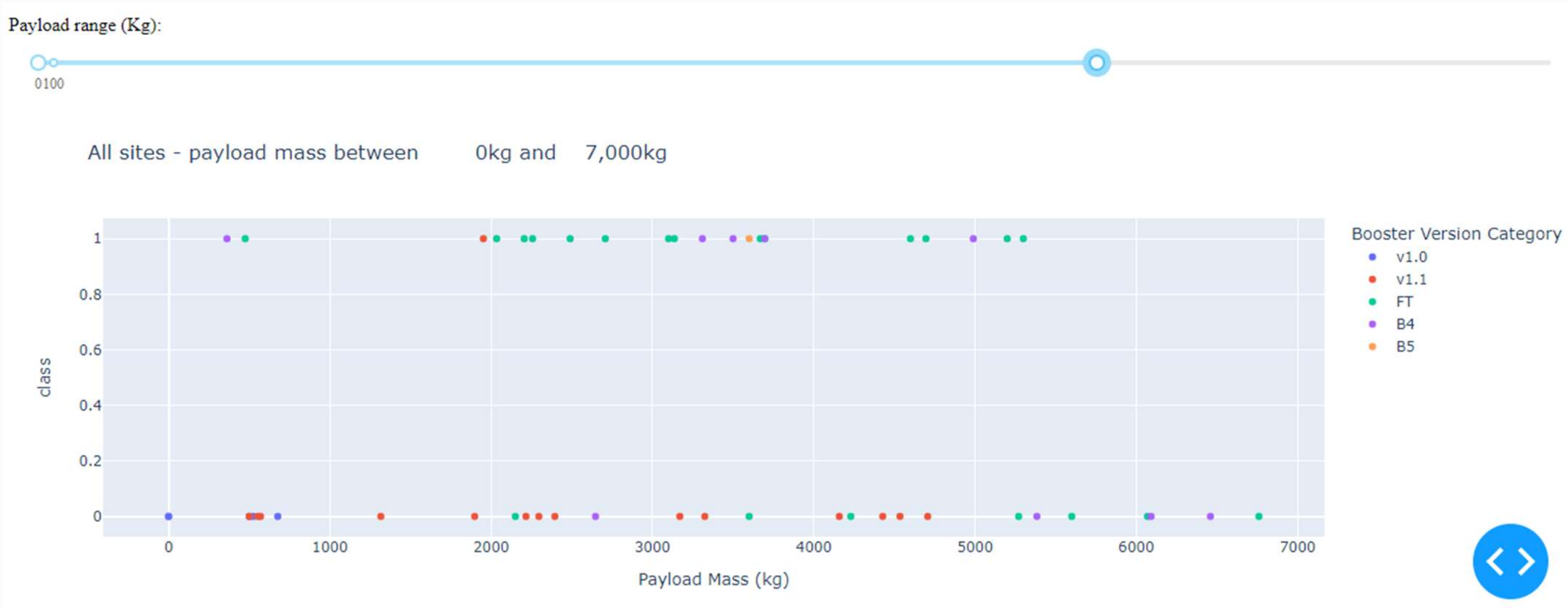
Build a Dashboard with Plotly Dash

Successful Launches by Site



CCAFS LC-40 is the site with most success rate

Payload vs. Launch Outcome



Until 7,000 kg, FT is the most effective Booster Version Type, while v1.1 is the lesser of

Section 5

Predictive Analysis (Classification)

Classification Accuracy

- 4 classification models were tested using the test set
  - Logistic Regression
  - SVM
  - Tree Decision
  - KNN

	LogisticRegression	SVM	Decision Tree	KNN
Jaccard_Score	0.800000	0.800000	0.800000	0.800000
F1_Score	0.888889	0.888889	0.888889	0.888889
Accuracy	0.833333	0.833333	0.833333	0.833333

- All models were trained to find out the best hyperparameters in the training set. The same model optimized was applied in the test set to evaluate accuracy
- As the results shown, all the models had the same performance, probably because the test set was too small, with only 18 samples

Classification Accuracy

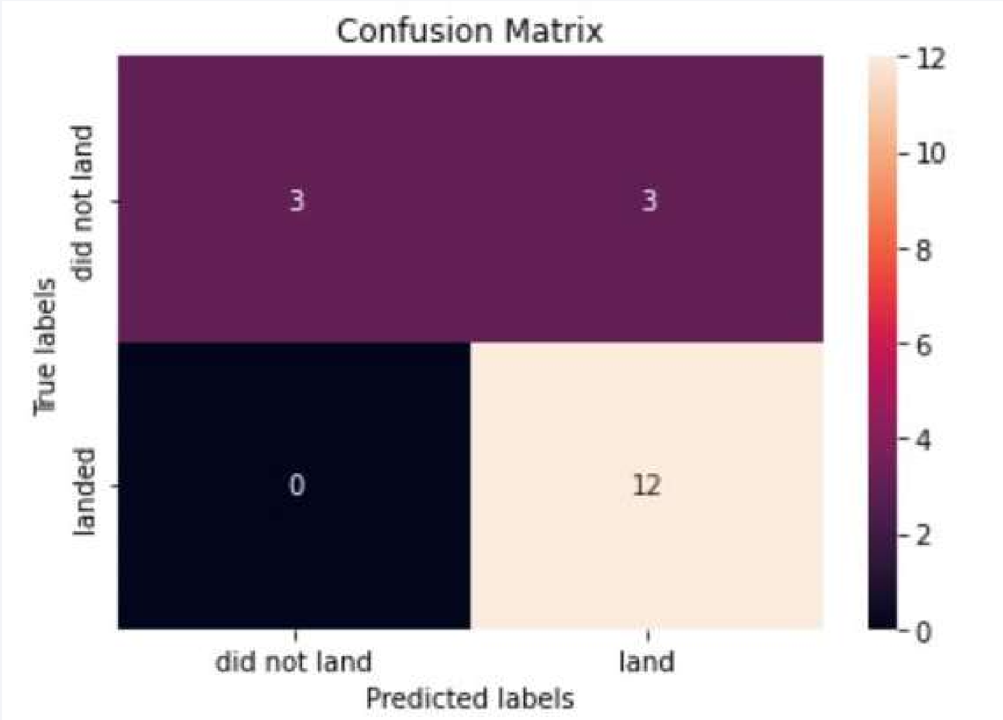
- However, when we tested the models with the whole dataset, we got different results, as it follows

	LogisticRegression	SVM	Decision Tree	KNN
Jaccard_Score	0.833333	0.845070	0.779221	0.819444
F1_Score	0.909091	0.916031	0.875912	0.900763
Accuracy	0.866667	0.877778	0.811111	0.855556

- So, as we can see, the SVM model had the best performance, not only on Accuracy, but on F1 Score and Jaccard Score too

Confusion Matrix

As said before, all models had the same performance with the test set, so the confusion matrix for all of them is the same. So, the biggest problem is in the classification of "did not land", where half of the examples were misclassified.



Conclusions

- EDA (graphical and SQL) was crucial to evaluate the influence of features on successful landing (e.g. Launch site, orbit and payload mass)
  - Success rate have improved over time
  - Insights obtained from dashboards and geographic visualization
  - Most sites are close to the Equator line. All the sites are in very close to the coast.
  - Orbits ES-L1, GEO, HEO and SSO have 100% success rate.
- SVM classification model was the one with the best accuracy to preview the success of a rocket landing
  - Over 83% of accuracy on test dataset
  - Over 87% of accuracy on full dataset


Thank you!